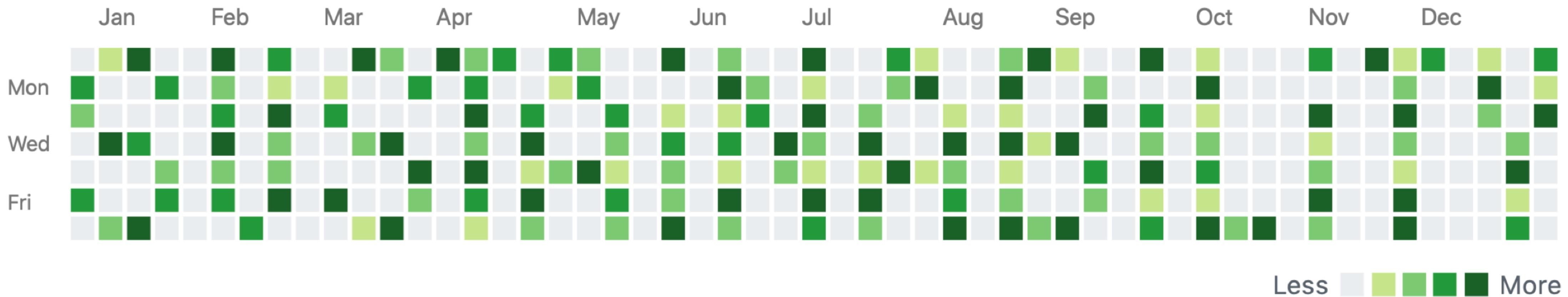


What can analyzing tens of terabytes of public trace data tell us about open source



Bogdan Vasilescu
@b_vasilescu

2020 top 10 “regions
outside the US by
open source use
(clones and
forks)”

<https://octoverse.github.com>



93% of Brazilian
federal gov't orgs
use open
source.

<https://www.zdnet.com/article/brazilian-federal-government-leads-in-open-source-adoption/>
(survey of 1686 federal and state-level orgs by
CGI.br between Jul-Oct 2017)



85% of Brazilian
federal gov't orgs
develop open
source.

<https://www.zdnet.com/article/brazilian-federal-government-leads-in-open-source-adoption/>
(survey of 1686 federal and state-level orgs by CGI.br between Jul-Oct 2017)

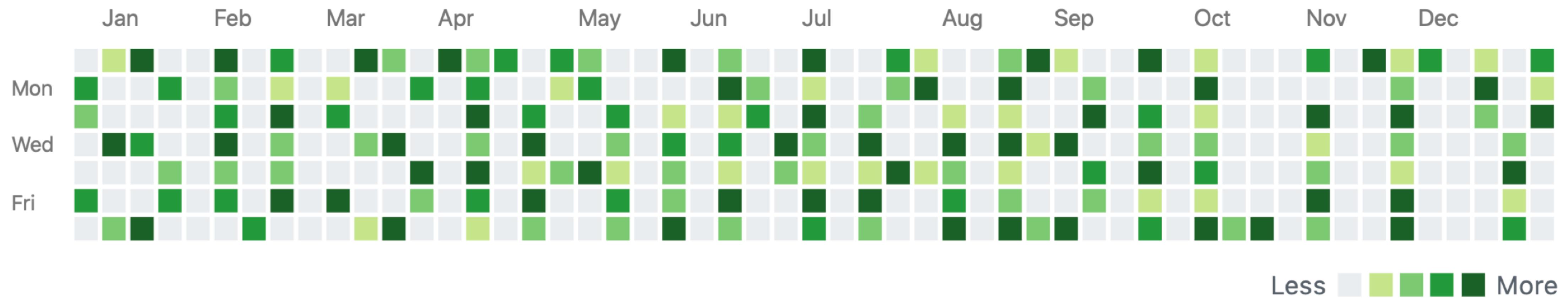


Sustaining
open source
is hard

However,

The fact that (almost) everything
is archived and public makes it
possible to study the problem
empirically

What can analyzing tens of terabytes of public trace data tell us about open source



This talk is about some of the things I learned

How I see
open source
today

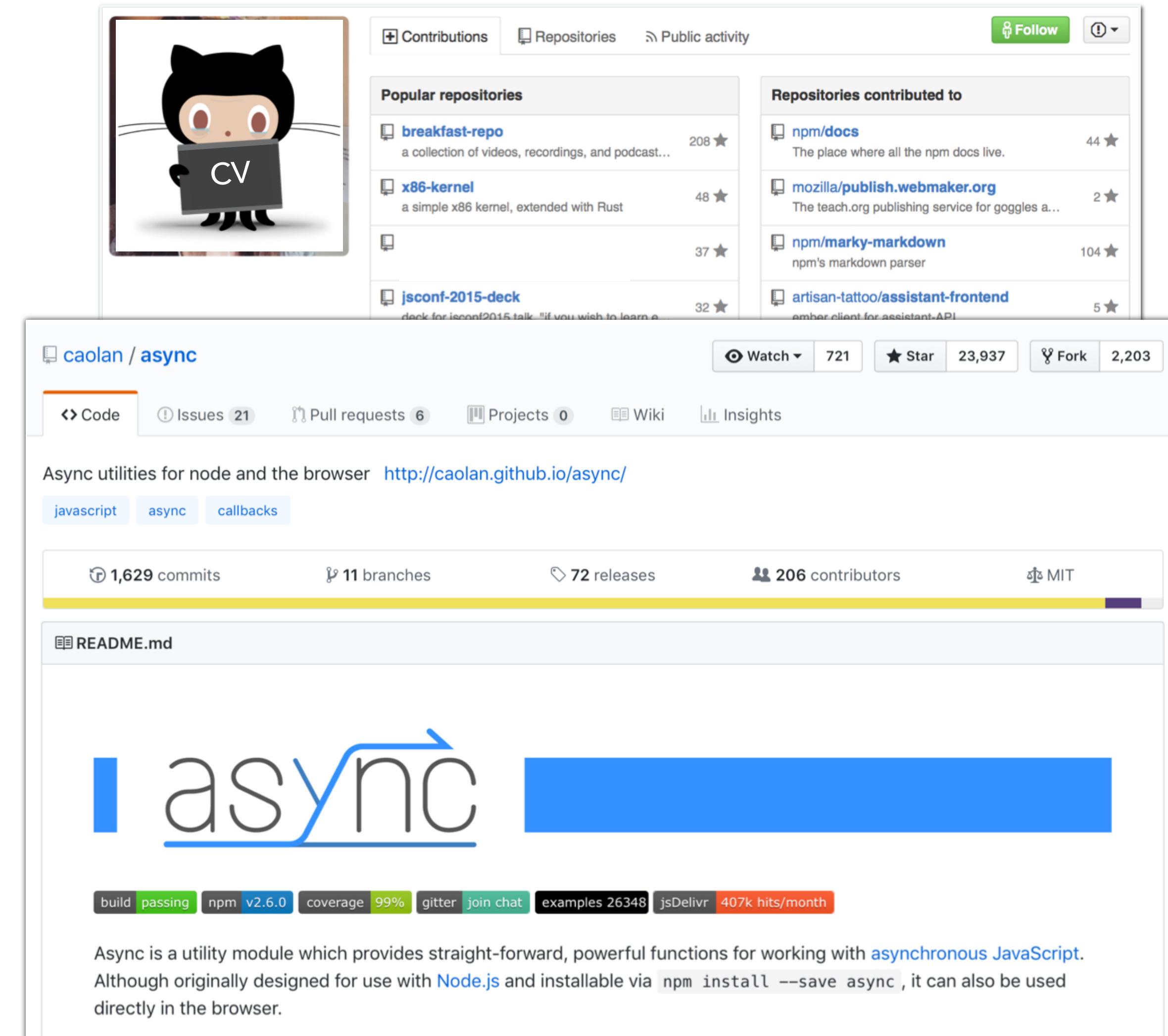
Change #1: More open source now than ever before

- Explosion of production in the past eight years

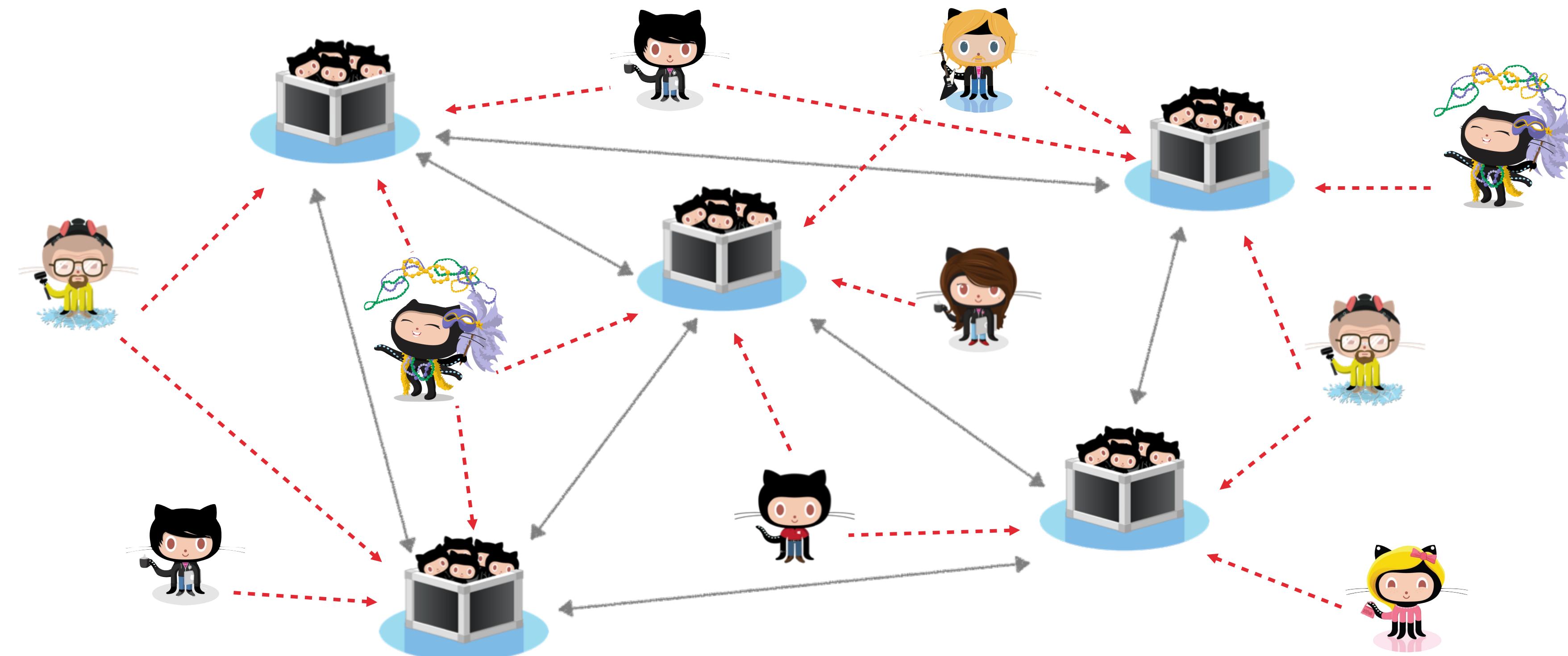


Change #2: The rise of social platforms

- Profile pages for users and projects
- 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)



Change #3: Complex socio-technical ecosystems



Interconnections & dependencies

Can be brittle

The Heartbleed Bug

The Heartbleed Bug is a serious vulnerability in the popular OpenSSL cryptographic software library. This weakness allows stealing the information protected, under normal conditions, by the SSL/TLS encryption used to secure the Internet. SSL/TLS provides communication security and privacy over the Internet for applications such as web, email, instant messaging (IM) and some virtual private networks (VPNs).

The Heartbleed bug allows anyone on the Internet to read the memory of the systems protected by the vulnerable versions of the OpenSSL software. This compromises the secret keys used to identify the service providers and to encrypt the traffic, the names and passwords of the users and the actual content. This allows attackers to eavesdrop on communications, steal data directly from the services and users and to impersonate services and users.



What leaks in practice?

We have tested some of our own services from attacker's perspective. We attacked ourselves from outside, without leaving a trace. Without using any privileged information or credentials we were able steal from ourselves the secret keys used for our X.509 certificates, user names and passwords, instant messages, emails and business critical documents and communication.

How to stop the leak?

As long as the vulnerable version of OpenSSL is in use it can be abused. Fixed OpenSSL (<https://www.openssl.org/news/secadv/20140407.txt>) has been released and now it has to be deployed. Operating system vendors and distribution, appliance vendors, independent software vendors have to adopt the fix and notify their users. Service providers and users have to install the fix as it becomes available for the operating systems, networked appliances and software they use.

<https://heartbleed.com>

NPM ERR!

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

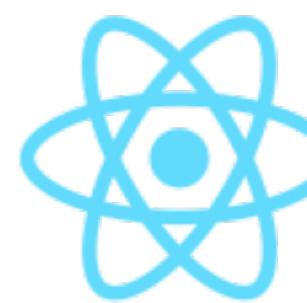
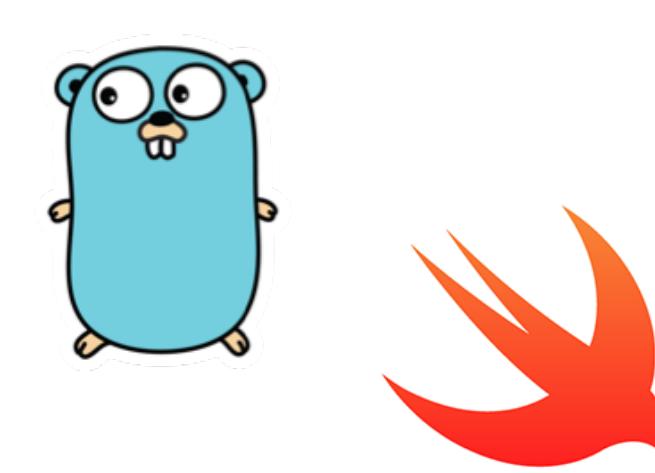
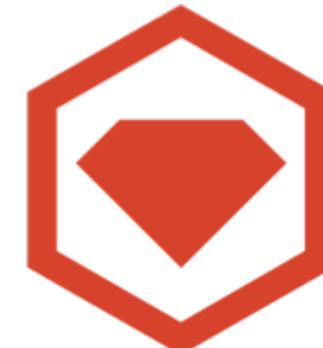
By Keith Collins • March 27, 2016

```
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/>

Change #4: Increasing commercialization and professionalization

- Historically
 - Community-based projects
(Python, RubyGems, Twisted)
- Currently
 - Lots of commercial involvement
 - Companies (Go - Google, React - Facebook, Swift - Apple)
 - Startups (Docker, npm, Meteor)



- 23% of respondents to 2017 GitHub survey:
job duties include contributing to open source

<http://opensourcesurvey.org/2017/>

Change #5: High expectations toward the quality, reliability, and security of open source infrastructure

- 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.



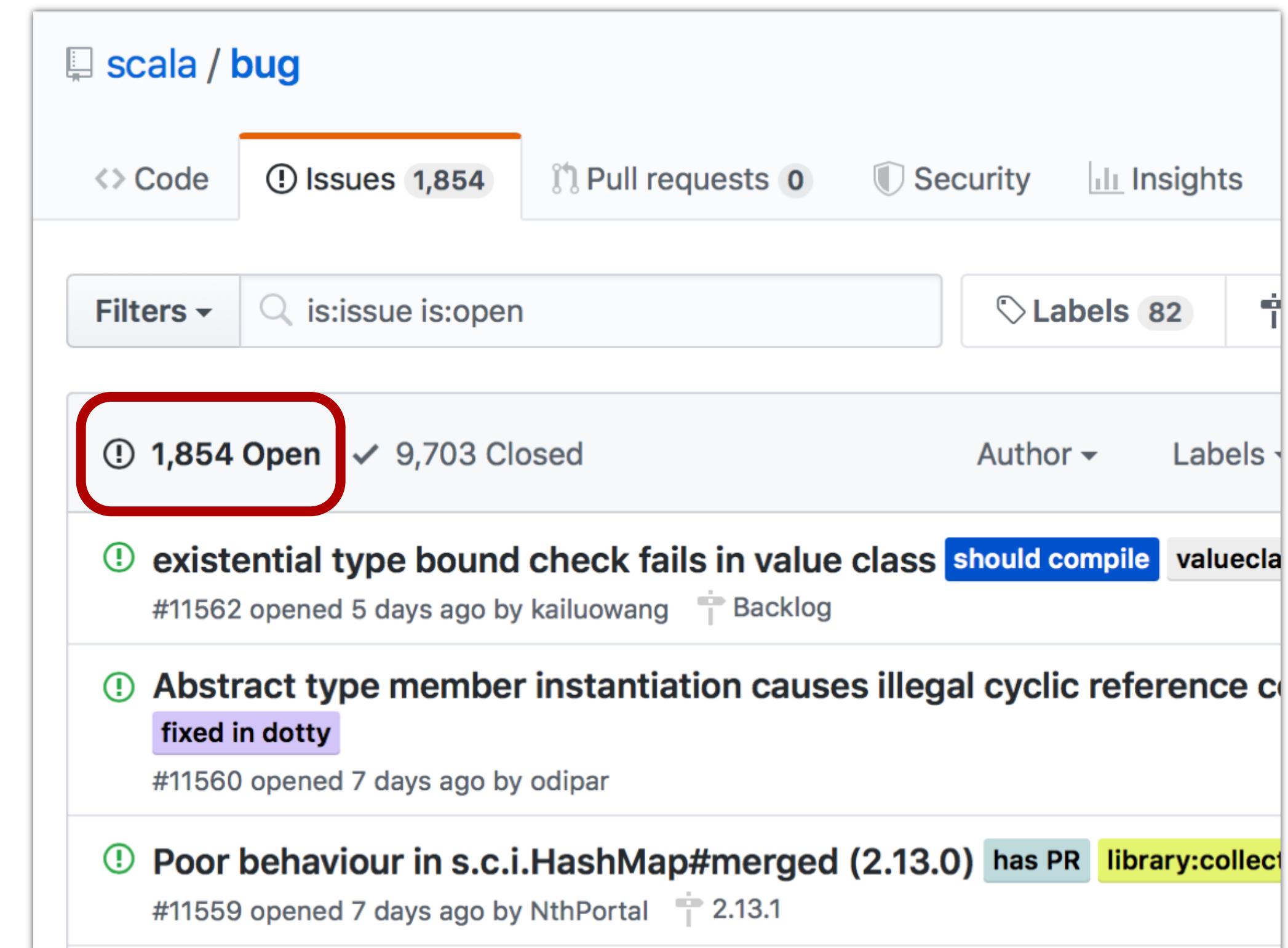
By Zack Whittaker | September 14, 2017 -- 01:27 GMT (18:27 PDT) | Topic: Security



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

Change #6: High level of demands & stress

- Easy to report issues / submit PRs
 - Growing volume of requests
- 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.”*



Lots of change, lots of challenges

- Best practices?
- What works?
- What doesn't?
- Long term sustainability?
- Equitable and healthy interactions?

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

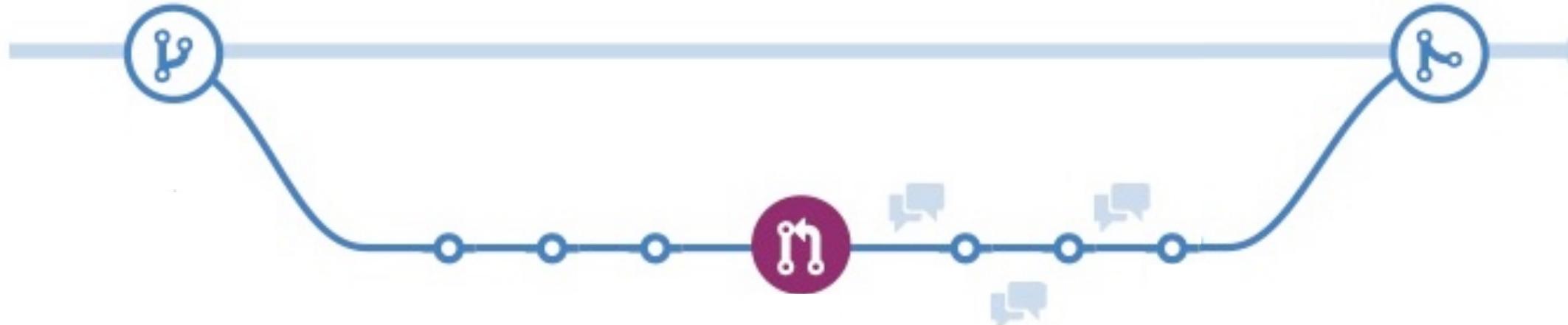
A great opportunity
for research

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 top navigation bar includes 'Overview', 'Repositories 23', 'Projects', and 'Packages'. Below the header is a circular profile picture of a man with glasses. The main content area starts with 'Popular repositories': 1. 'empirical-methods': Fall 2018 edition of 17-803 Empirical Methods, 12 stars, 1 fork. 2. 'diversity': A data set for social diversity studies of GitHub teams, 11 stars, 5 forks. 3. 'jsNaughty': JS reverse minifier based on statistical machine translation, JavaScript, 8 stars, 2 forks. 4. 'ght_unmasking_aliases': Python, 6 stars, 6 forks. 5. 'bvasiles.github.io': My website, HTML, 2 stars. 6. 'SuffixTree': Forked from JDonner/SuffixTree, C, 1 star. The profile summary shows 40 followers, 5 following, and 27 public repos. It also lists Carnegie Mellon University as the affiliation and Pittsburgh, PA as the location, with links to the website and Twitter handle (@b_vasilescu).

→ Uniform access to contribution and personal data

Heaps of data



GitHub alone:

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

W $\ltimes\gt;$ C Beyond GitHub:

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

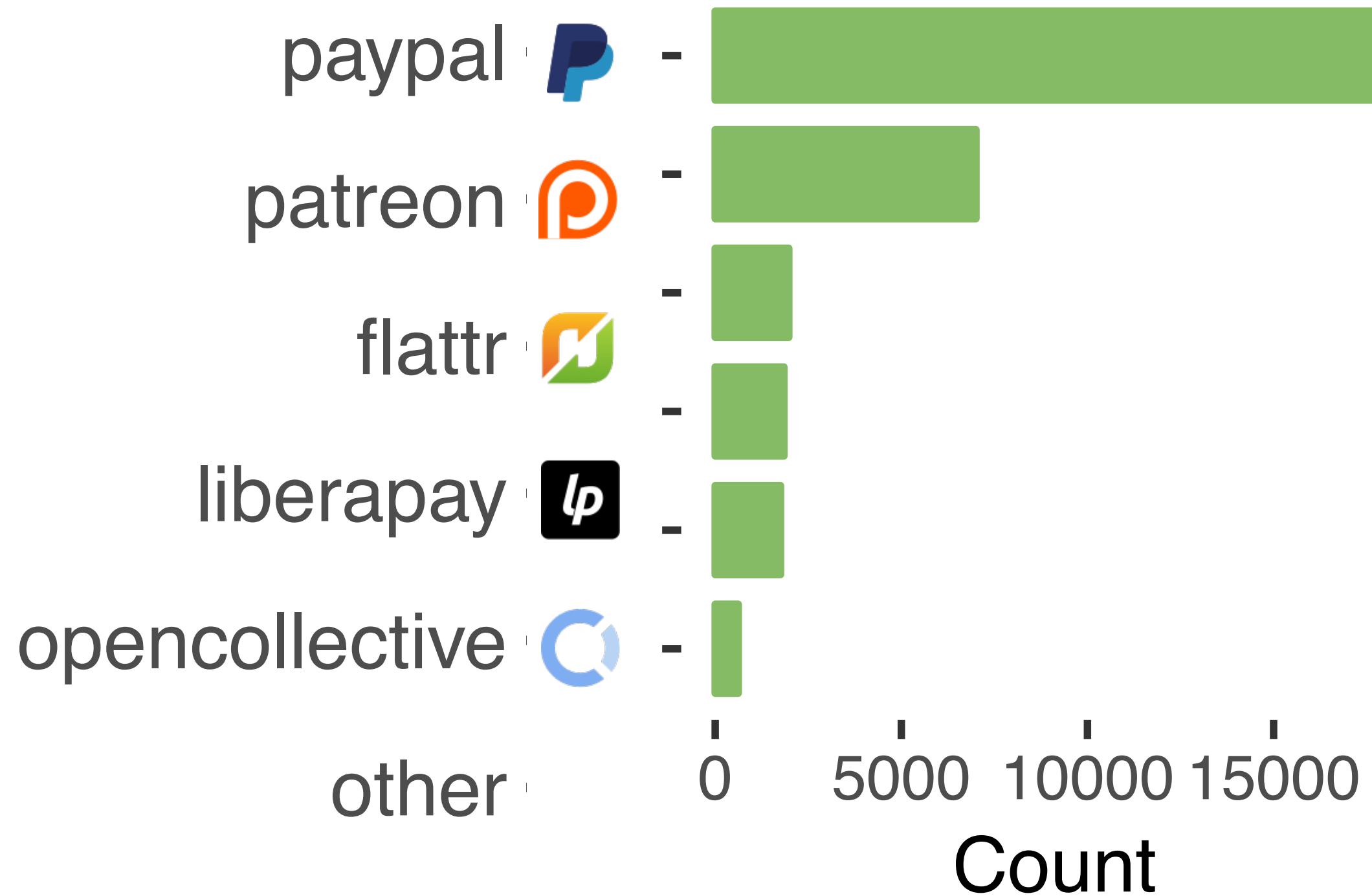
Ma, Y., Bogart, C., Amreen, S., Zaretsky, R., & Mockus, A. (2019, May). World of Code: An infrastructure for mining the universe of open source VCS data. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)* (pp. 143-154). IEEE.

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



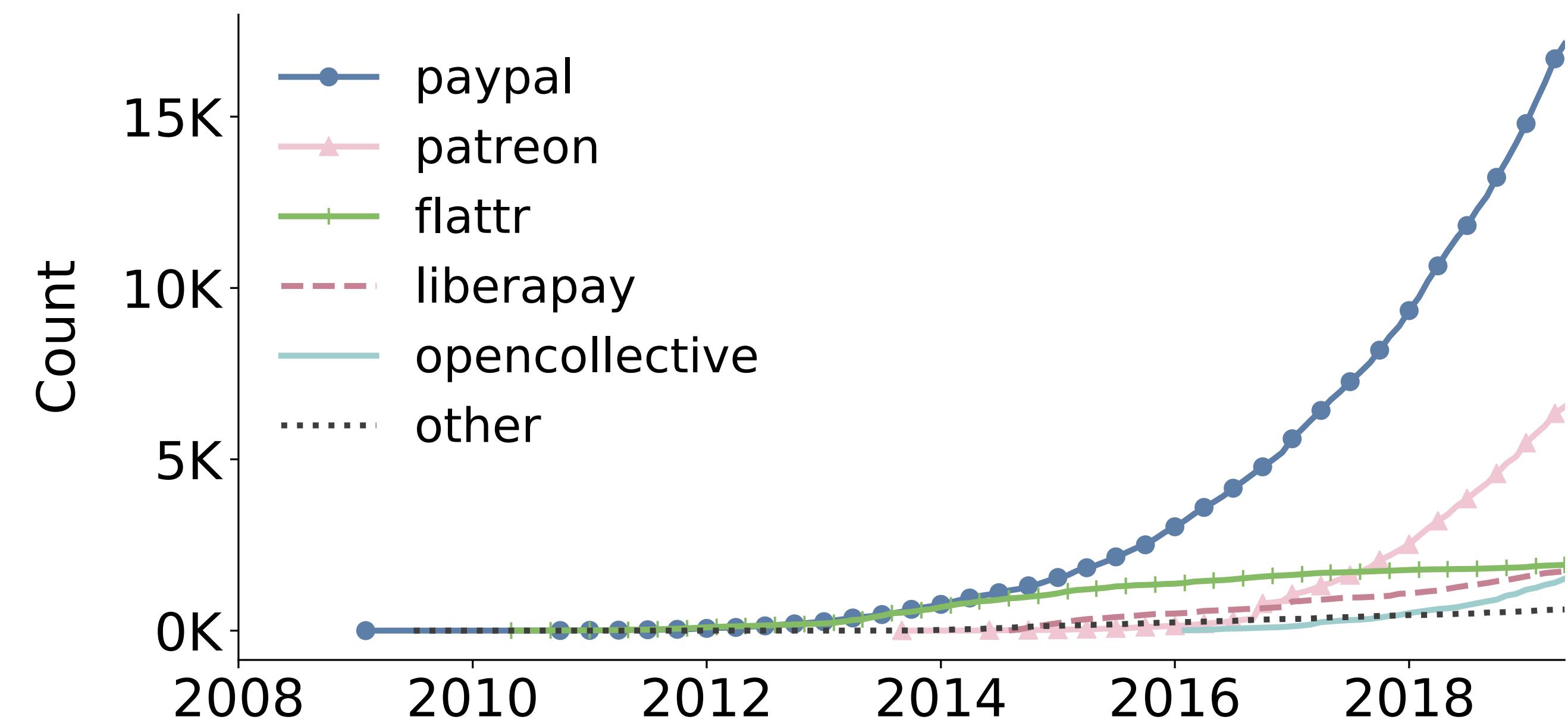
as of May 23, 2019

A great opportunity for research

Capture and understand trends over time, analyze time series data

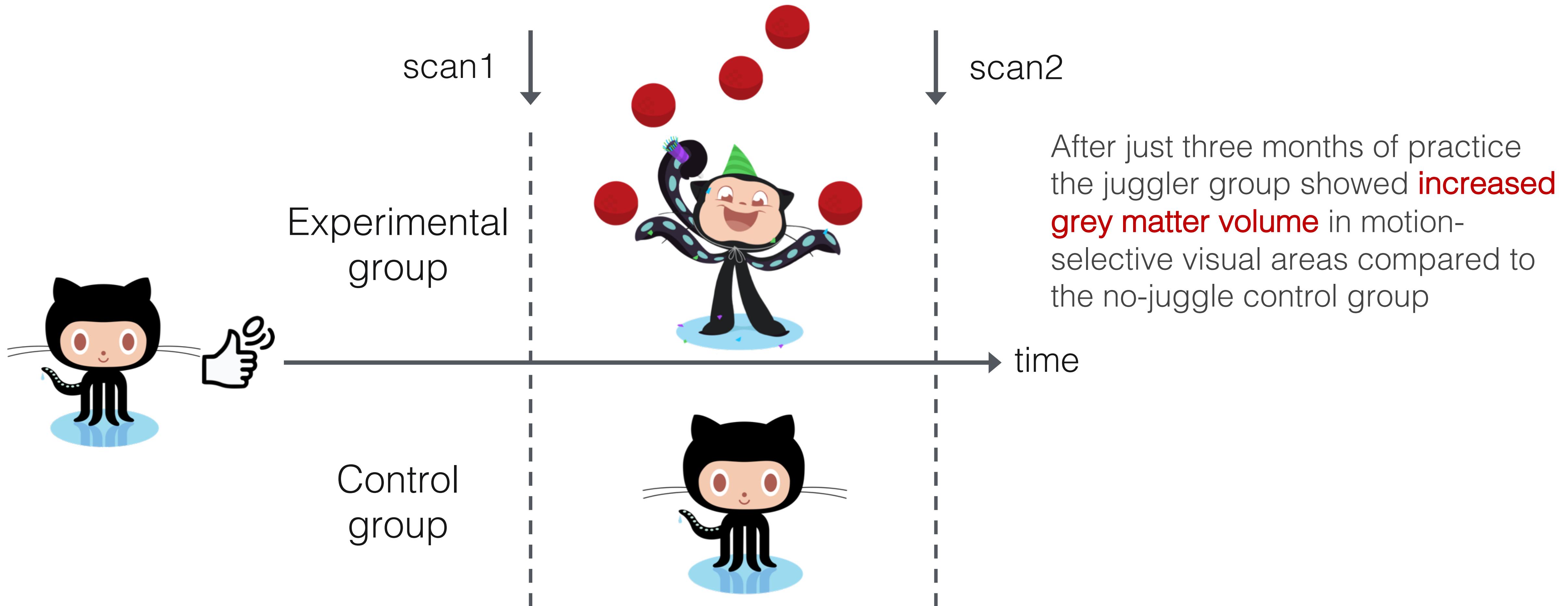


Adoption of donation platforms over time



Aside: Experiment

Juggling as a sustainability intervention?

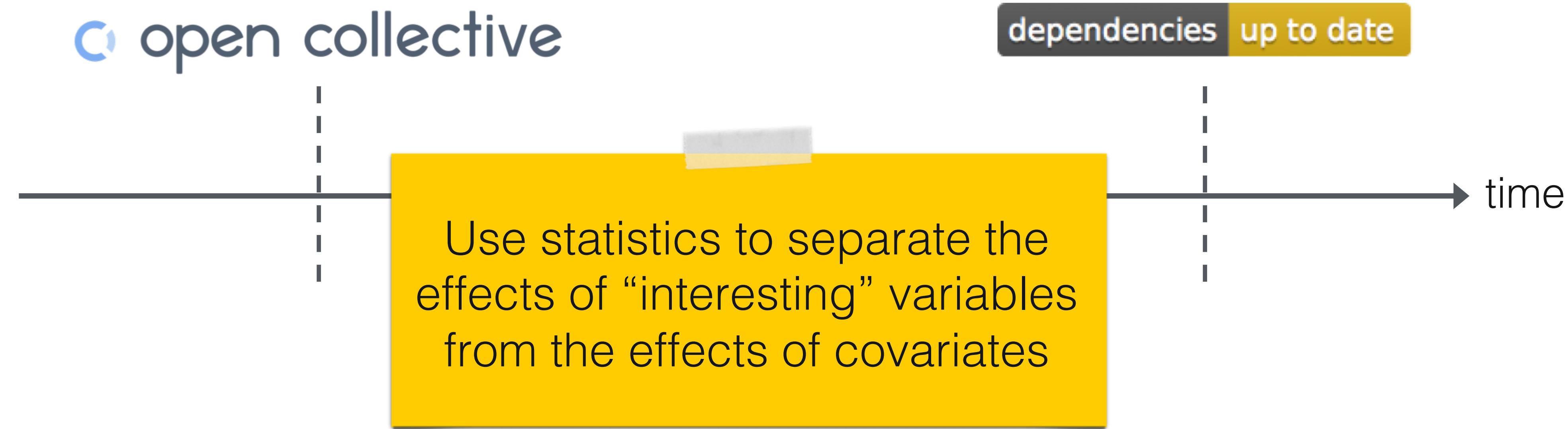


Bogdan Draganski, Christian Gaser, V. Busch, G. Schuierer, U. Bogdahn, and A. May. "Changes in grey matter induced by training." Nature 427, no. 6972 (2004): 311-312.

Natural experiments: interventions are outside the control of the researchers

Project started receiving donations

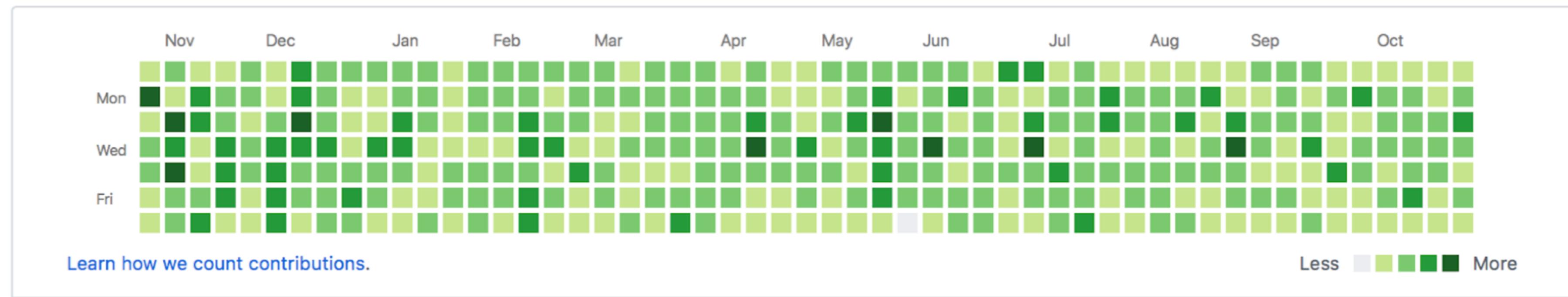
Project adopted a certain practice / tool



But be careful, the data is noisy!

- Hyperactive maintainer? No, bot

5,786 contributions in the last year



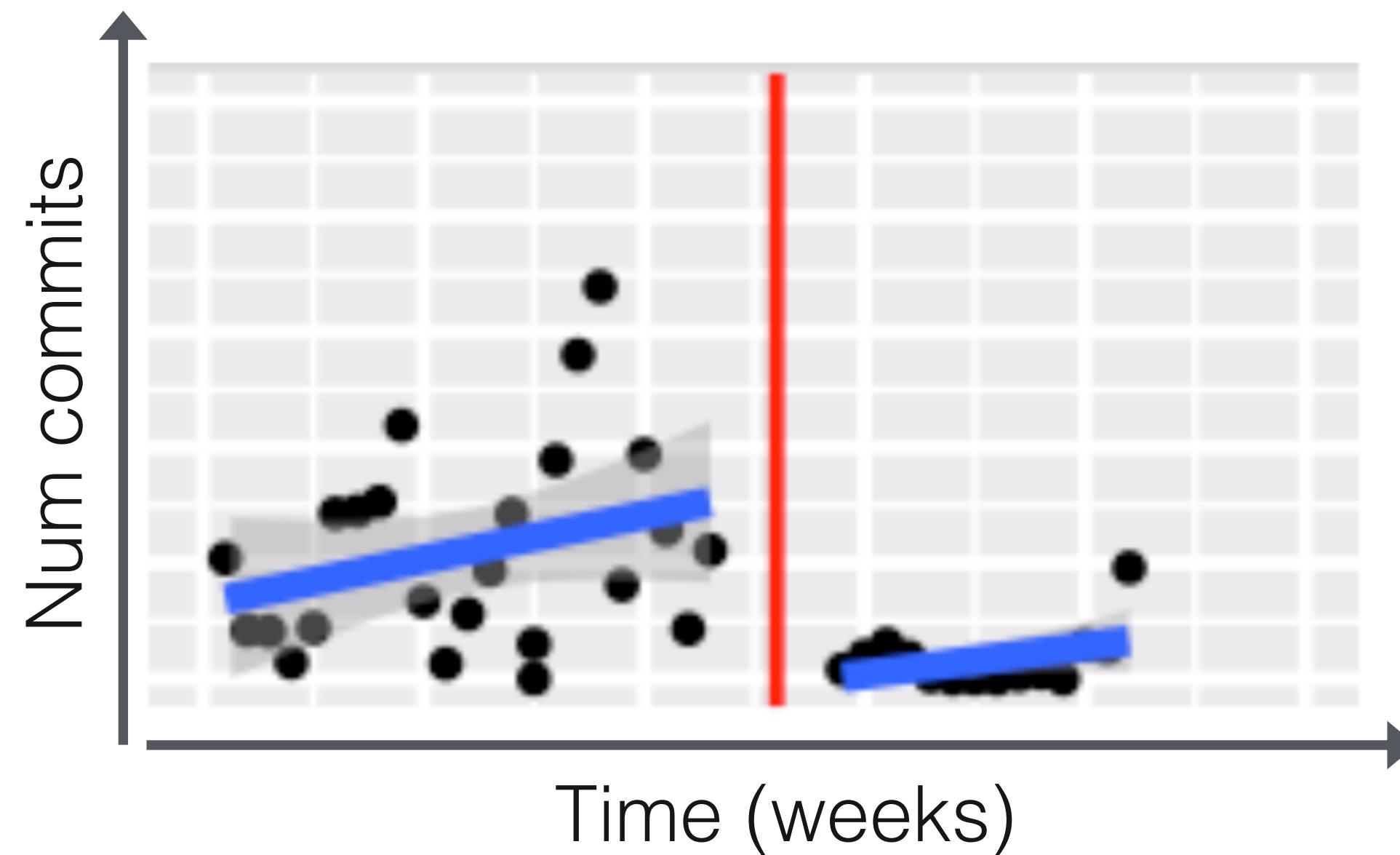
fossabot

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!

Why did this person drop out?



- Social science theory
- Qualitative analysis (surveys, interviews)

Journal of Applied Psychology
2017, Vol. 102, No. 3, 530–545

© 2017 American Psychological Association
0021-9010/17/\$12.00 <http://dx.doi.org/10.1037/apl0000103>

One Hundred Years of Employee Turnover Theory and Research

Peter W. Hom
Arizona State University

Thomas W. Lee
University of Washington

Jason D. Shaw
Hong Kong Polytechnic University

John P. Hausknecht
Cornell University

We review seminal publications on employee turnover during the 100-year existence of the *Journal of Applied Psychology*. Along with classic articles from this journal, we expand our review to include other publications that yielded key theoretical and methodological contributions to the turnover literature. We first describe how the earliest papers examined practical methods for turnover reduction or control and then explain how theory development and testing began in the mid-20th century and dominated the academic literature until the turn of the century. We then track 21st century interest in the psychology of staying (rather than leaving) and attitudinal trajectories in predicting turnover. Finally, we discuss the rising scholarship on collective turnover given the centrality of human capital flight to practitioners and to the field of human resource management strategy.

Let's look at some
concrete examples

1.

Open Source and money

A handy guide to financial support for open source.

"I do open source work, how do I find funding?"

This document aims to provide an exhaustive list of all the ways that people get paid for open source work. Hopefully, projects and contributors will find this helpful in figuring out the best options for them.

The list below is roughly ordered from small to large. Each funding category links to several real examples (using topical articles or pages wherever possible instead of just a project's homepage.)

The categories are not mutually exclusive. For example, a project might have a foundation but also use crowdfunding to raise money. Someone else might do consulting and also have a donation button. Etc.

Table of Contents

1. [Donation button](#)
2. [Bounties](#)
3. [Sponsorware](#)
4. [Crowdfunding \(one-time\)](#)
5. [Crowdfunding \(recurring\)](#)
6. [Books and merchandise](#)
7. [Advertising & sponsorships](#)

<https://github.com/nayafia/lemonade-stand>



Donations are gaining in popularity as a potential solution

Only anecdotes about their prevalence and impact

Caleb Porzio (@calebporzio)

I just cracked \$100k/yr on GitHub Sponsors. ❤️🎉
Life. Changed.
Huge thanks to everyone's generosity and the GH Sponsors team! ❤️🙏

Did a writeup of the entire journey if you care:
calebporzio.com/i-just-hit-dol...

TOTAL SPONSORS
535
[View all →](#)

MONTHLY ESTIMATED INCOME
\$9,390.00
This is an estimate of monthly income based on current monthly and yearly sponsorships.

112680
C % ÷
7 8 9 ×
4 5 6 –
1 2 3 +
0 . =

Your GitHub Sponsors profile
[Read more about managing your profile.](#)

Next steps
Here are some things you can do to grow your sponsorship

9:42 AM · Jun 23, 2020

5.2K 778 people are Tweeting about this

Chris Aniszczyk (@cra)

paying maintainers via charity or donations is the wrong approach for long term sustainability, also shorts maintainers into a gig-style economy without benefits, it's corporations that need to give back through hiring and setting time for open source contribution

sMyle (@MylesBorins)

Open source doesn't work without large scale enterprise or corporate investment

Simply paying maintainers has the wrong incentive model and is not scalable. [twitter.com/AmarachiAmaech...](#)

8:07 PM · Mar 10, 2019 from San Carlos, CA

57 See Chris Aniszczyk's other Tweets

Lots to explore...

GitHub-scale
census of donation
requests

Stated
expectations
for donations

Actual
usage
of donations

Characteristics
of projects asking /
getting money

Measurable
effects
of donations

How to Not Get Rich: An Empirical Study of Donations in Open Source.
Overney, C., Meinicke, J., Kästner, C., and Vasilescu, B. *ICSE*, ACM (2020).

Key insight for identifying donation platforms: README files contain signals of donation requests

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

Babel (pronounced "babble") is a community-driven project used by many companies and projects, and is maintained by a group of [volunteers](#). If you'd like to help support the future of the project, please consider:

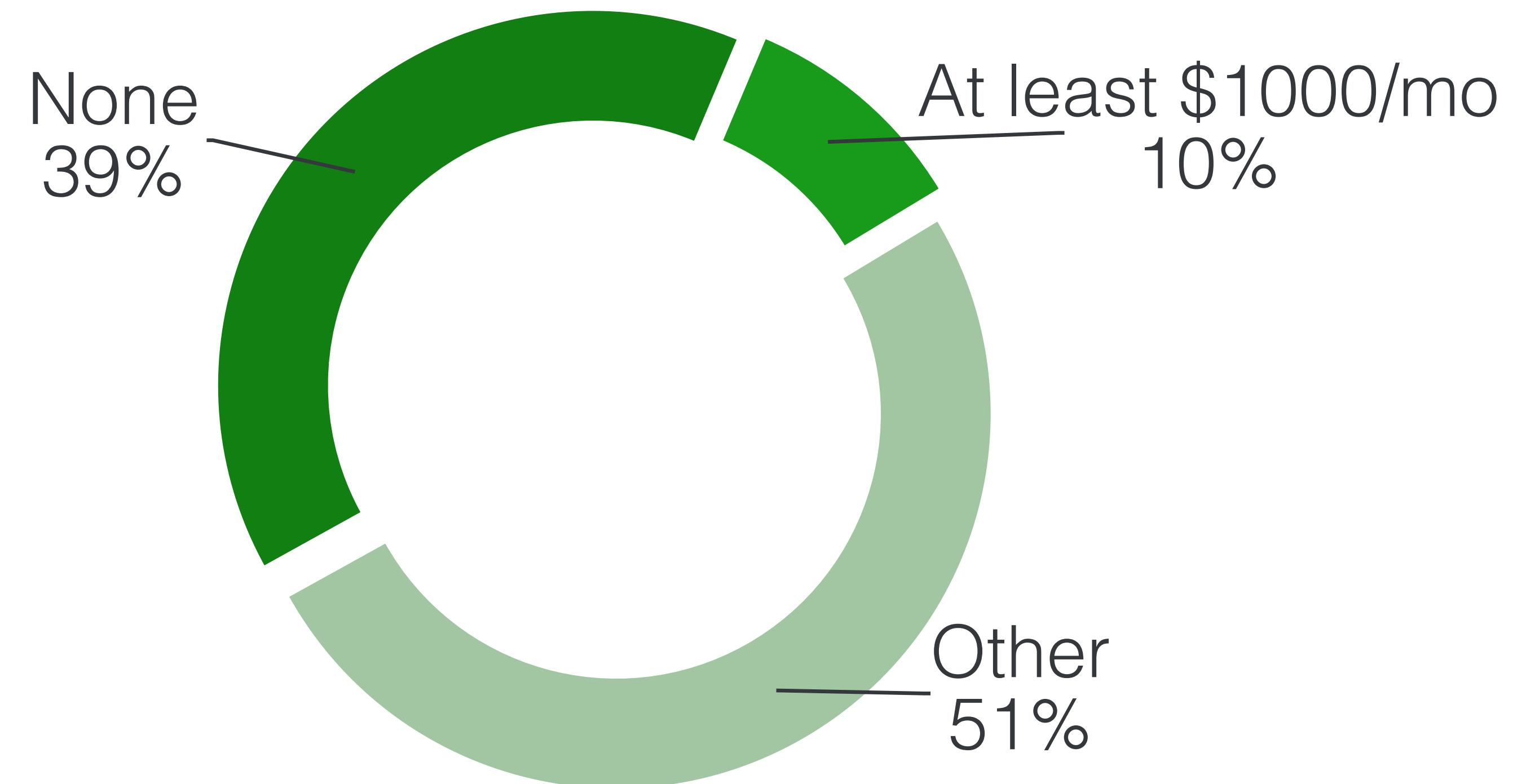
- Giving developer time on the project. (Message us on [Twitter](#) or [Slack](#) for guidance!)
- Giving funds by becoming a sponsor on [Open Collective](#) or [Patreon](#)!

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

Most projects receive little funding

Sample: 6,516 repos using  patreon /  open collective

Census
Characteristics
Expectations
Effects
Usage



last 9 months before May 23, 2019

Statistical multi-variate analysis

	<i>Resp: Asks for donations</i>	
	Coeffs (Err.)	Deviance
(Intercept)	-4.01 (0.19)***	
commits (log)	0.40 (0.05)***	72.95***
size (log)	-0.30 (0.03)***	125.74***
project age	0.02 (0.00)***	85.94***
is active	1.95 (0.09)***	502.20***
is org	-0.57 (0.10)***	33.63***
stars (log)	0.27 (0.02)***	129.89***
downloads (log)	-0.02 (0.02)	0.88
dependents (log)	0.01 (0.05)	0.02
Num. obs.	9137	
R ²	0.31	

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

	Hurdle model		Count model	
	<i>Resp: Received any</i>		<i>Resp: Amount received</i>	
	Coeffs (Err.)	Deviance	Coeffs (Err.)	Sum sq.
(Intercept)	0.12 (0.38)		4.17 (0.39)***	
commits (log)	-0.20 (0.12)	3.05	-0.26 (0.11)*	20.41*
size (log)	-0.10 (0.06)	2.80	0.06 (0.07)	2.67
project age	0.05 (0.01)***	58.63***	-0.01 (0.00)	10.93
is active	1.33 (0.22)***	38.73***	0.00 (0.21)	0.00
is org	0.84 (0.26)**	10.78**	0.12 (0.20)	1.37
stars (log)	0.14 (0.06)*	6.06*	0.39 (0.06)***	182.17***
downloads (log)	-0.11 (0.06)	3.51	0.13 (0.05)**	28.60*
dependents (log)	0.31 (0.11)**	8.98**	-0.04 (0.08)	0.85
Num. obs.	735		527	
R ²	0.29		0.30	

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

Census

Characterist.

Expectations

Effects

Usage

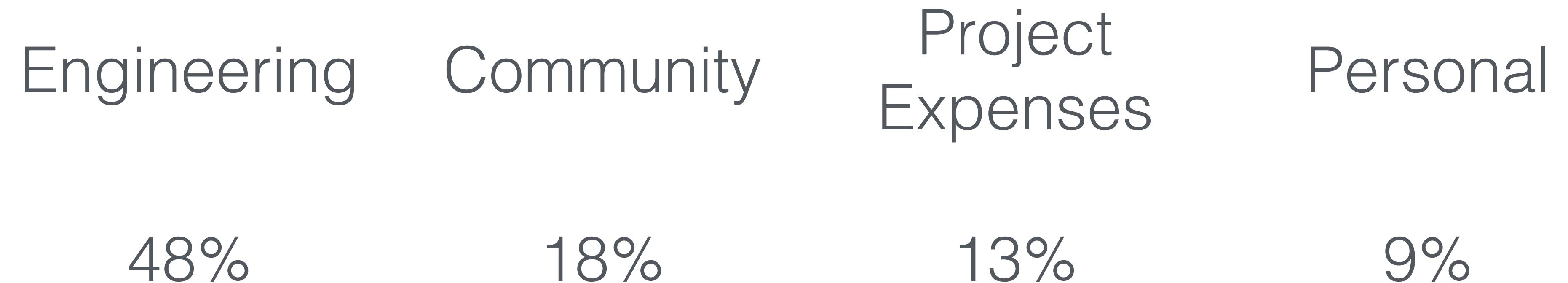
Projects receiving more donations...

more stars

more
downloads

Developers plan to spend donations on ...

Census
Characteristics
Expectations
Effects
Usage

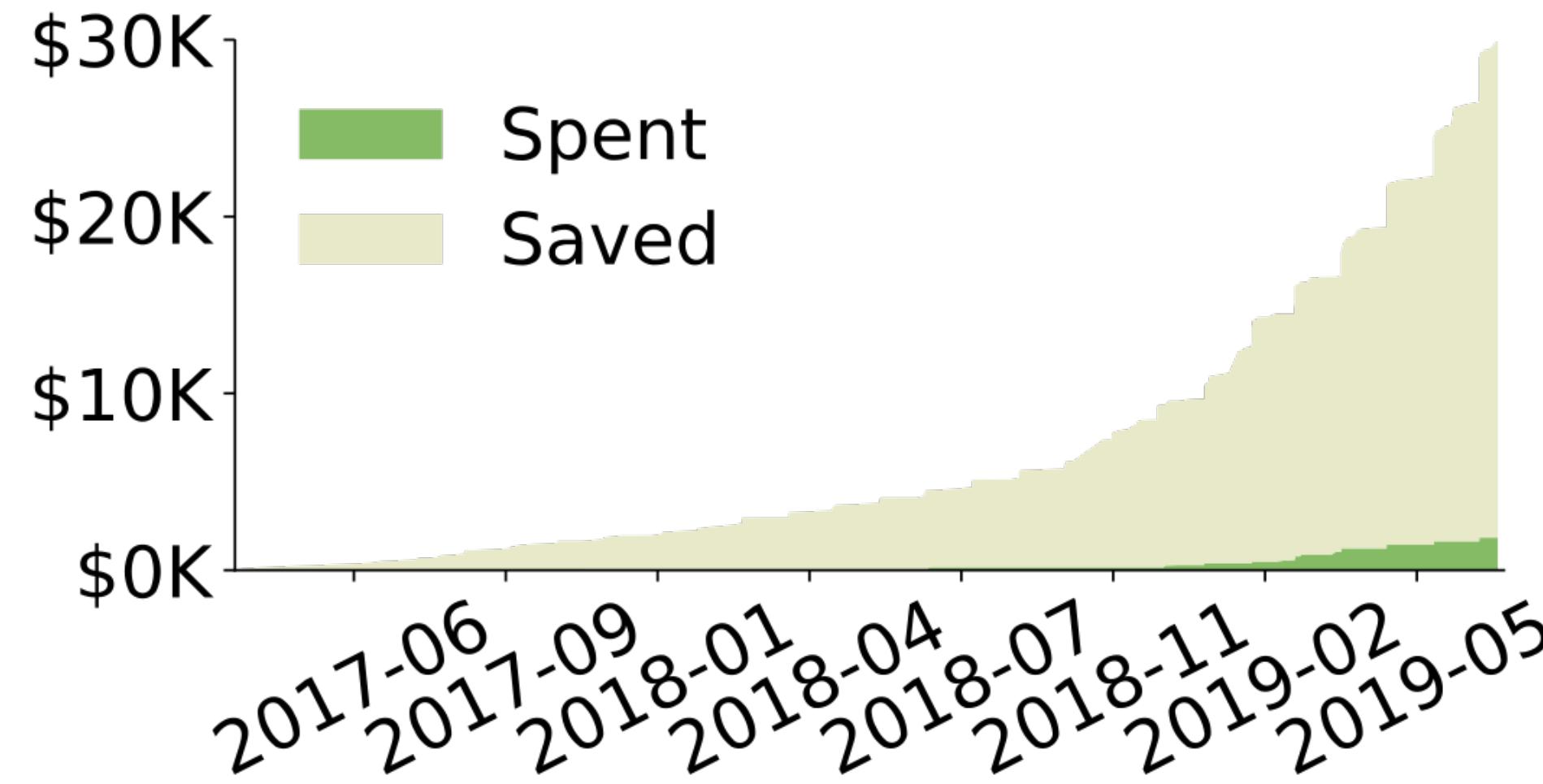


Qualitative analysis of donation profile pages for 109 **npm** projects on **patreon** **open collective** **KICKSTARTER**

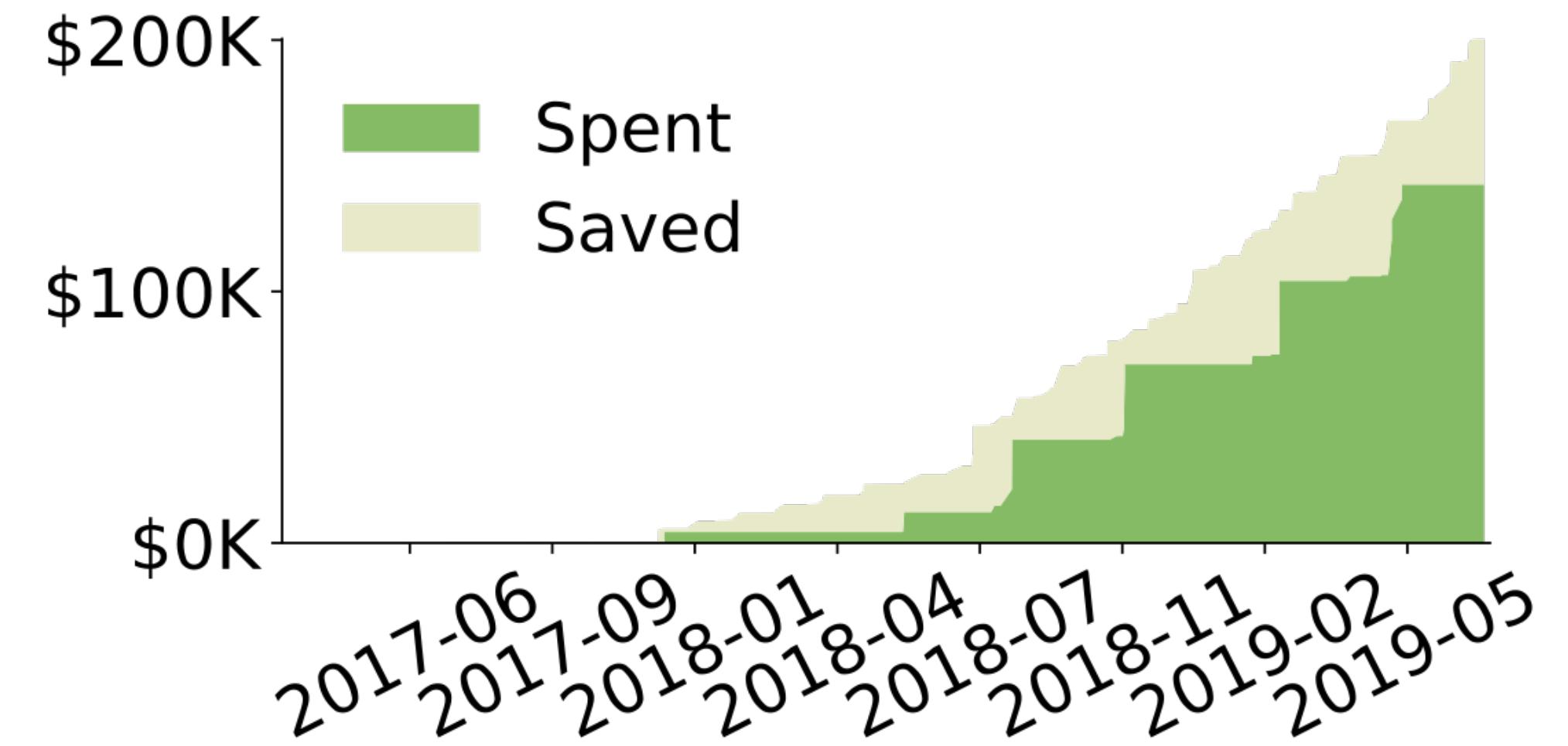
The use of donations varies widely: Savers vs spenders

Census
Characteristics
Expectations
Effects
Usage

64% Savers
spend less than 25% of raised
donations



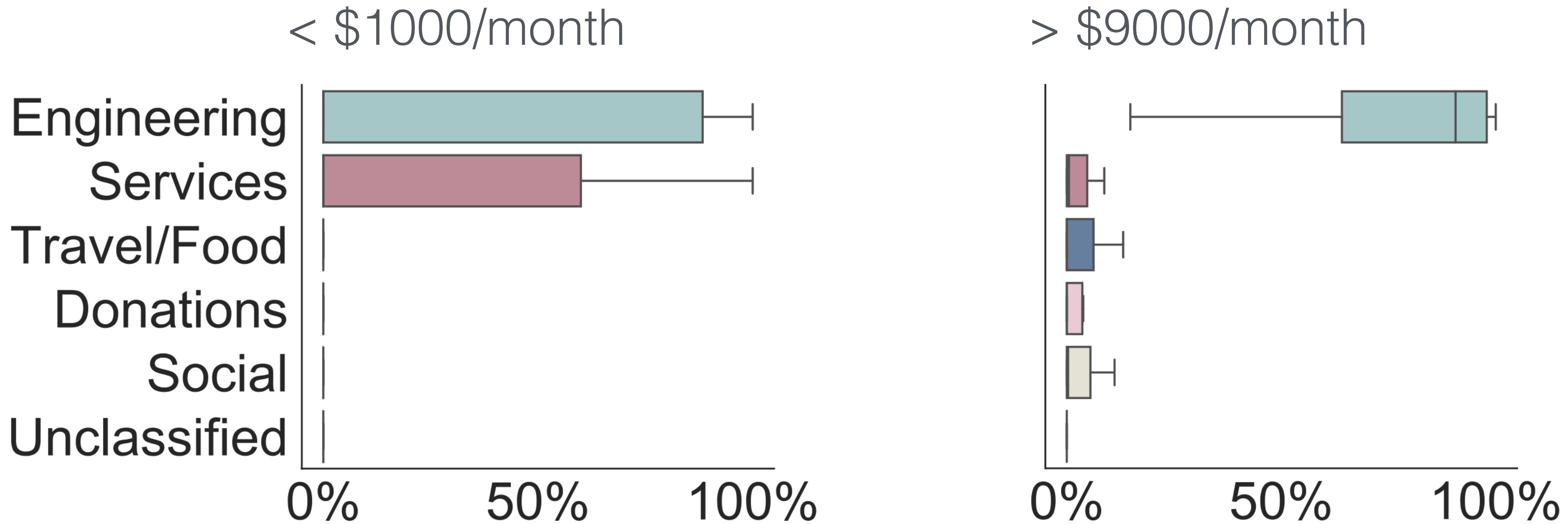
11% Spenders
spend more than 75% of raised
donations



The use of donations varies widely:

Type of expenses

Census
Characteristics
Expectations
Effects
Usage



Takeaways on how to effectively raise donations

Reputation
matters

Awareness
of need

Efficiency
of using
funds

Dark Side of
donations

Theory matters!

2. Transparency and signaling

Key insight for identifying donation platforms: README files contain signals of donation requests

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

Babel (pronounced "babble") is a community-driven project used by many companies and projects, and is maintained by a group of volunteers. If you'd like to help support the future of the project, please consider:

- Giving developer time on the project. (Message us on [Twitter](#) or [Slack](#) for guidance!)
- Giving funds by becoming a sponsor on [Open Collective](#) or [Patreon!](#)

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

Transparency is already a defining characteristic of the environment

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

- Popular repositories:**
 - breakfast-repo**: A collection of videos, recordings, and podcasts. 208 stars.
 - x86-kernel**: A simple x86 kernel, extended with Rust. 48 stars.
 - jsconf-2015-deck**: Deck for jsconf2015 talk, "if you wish to learn e...". 32 stars.
 - ratpack**: Sinatra boilerplate using activerecord, sqlite, a... 32 stars.
- Repositories contributed to:**
 - npm/docs**: The place where all the npm docs live. 44 stars.
 - mozilla/publish.webmaker.org**: The teach.org publishing service for goggles a... 2 stars.
 - npm/marky-markdown**: npm's markdown parser. 104 stars.
 - artisan-tattoo/assistant-frontend**: ember client for assistant-API. 5 stars.
 - npm/npm-camp**: A community conference for all things npm. 1 star.
- Public contributions:** A heatmap showing contributions by month and day over the last year.
- Contributions in the last year:** 1,886 total (Jan 24, 2015 – Jan 24, 2016).
- Longest streak:** 37 days (October 7 – November 12).
- Current streak:** 7 days (January 18 – January 24).
- Organizations:** Shows associations with npm, inc.
- Statistics:** 776 Followers, 38 Starred, 15 Following.

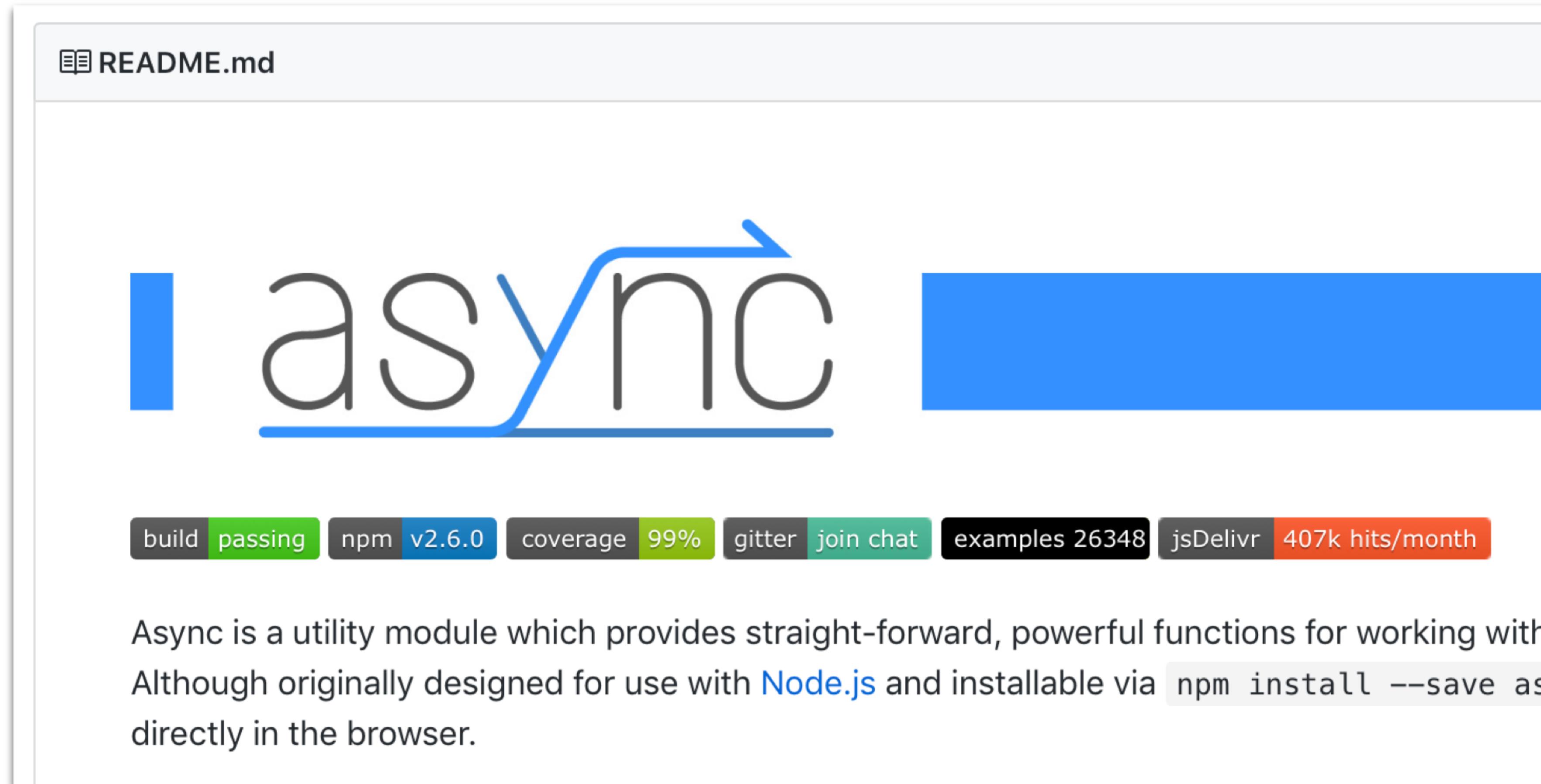
This screenshot shows the GitHub repository page for 'caolan/async'. The repository has 23,937 stars and 2,203 forks. The main sections include:

- Code**: Shows 1,629 commits, 11 branches, 72 releases, and 206 contributors.
- README.md**: Features a large 'async' logo with a blue arrow.
- Build status:** build passing.
- Dependencies:** npm v2.6.0, coverage 99%, gitter, join chat, examples 26348, jsDelivr, 407k hits/month.
- Description:** Async utilities for node and the browser. <http://caolan.github.io/async/>

Text at the bottom states: "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."

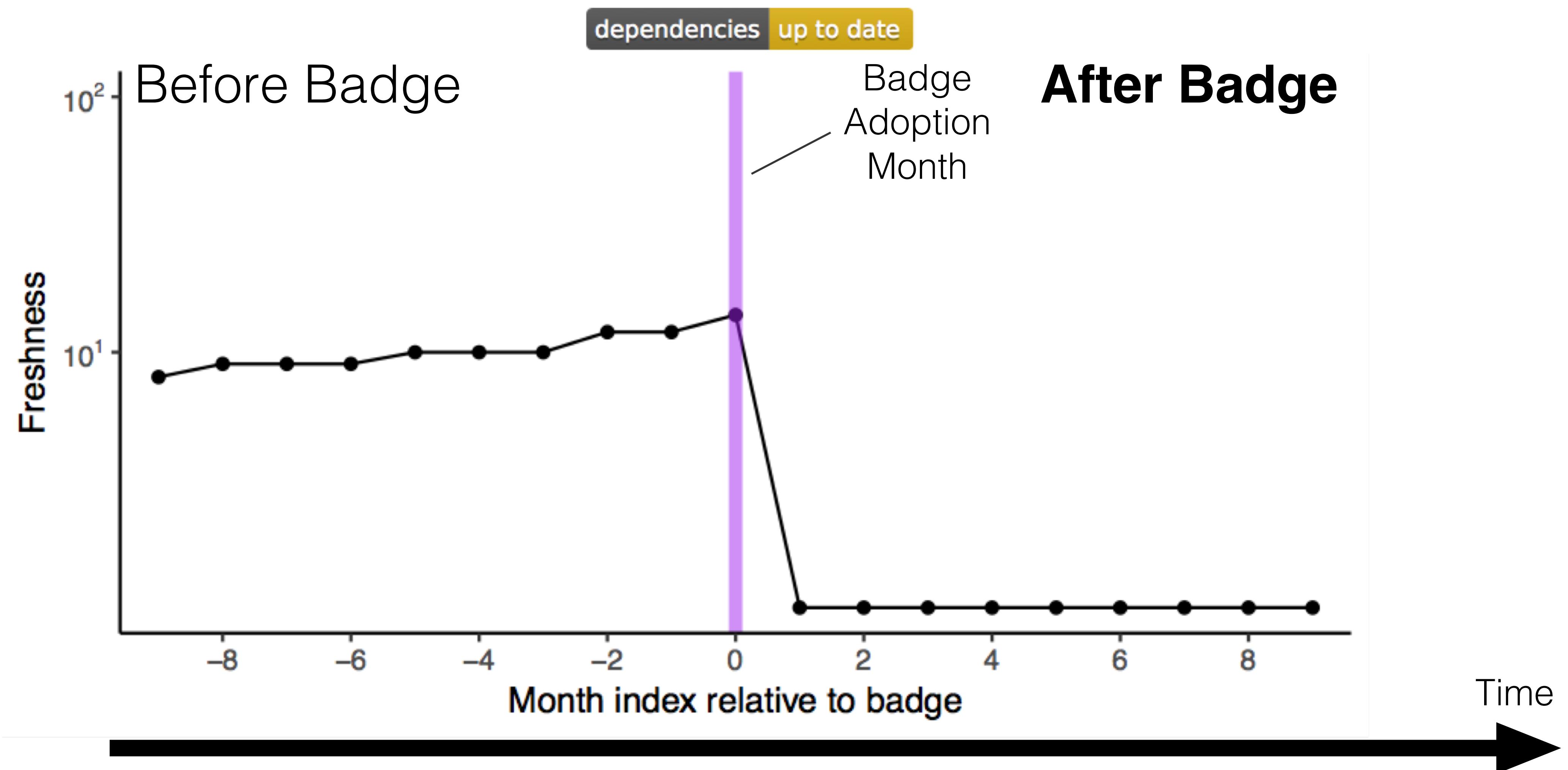
Signals are customizable

- E.g., repository badges

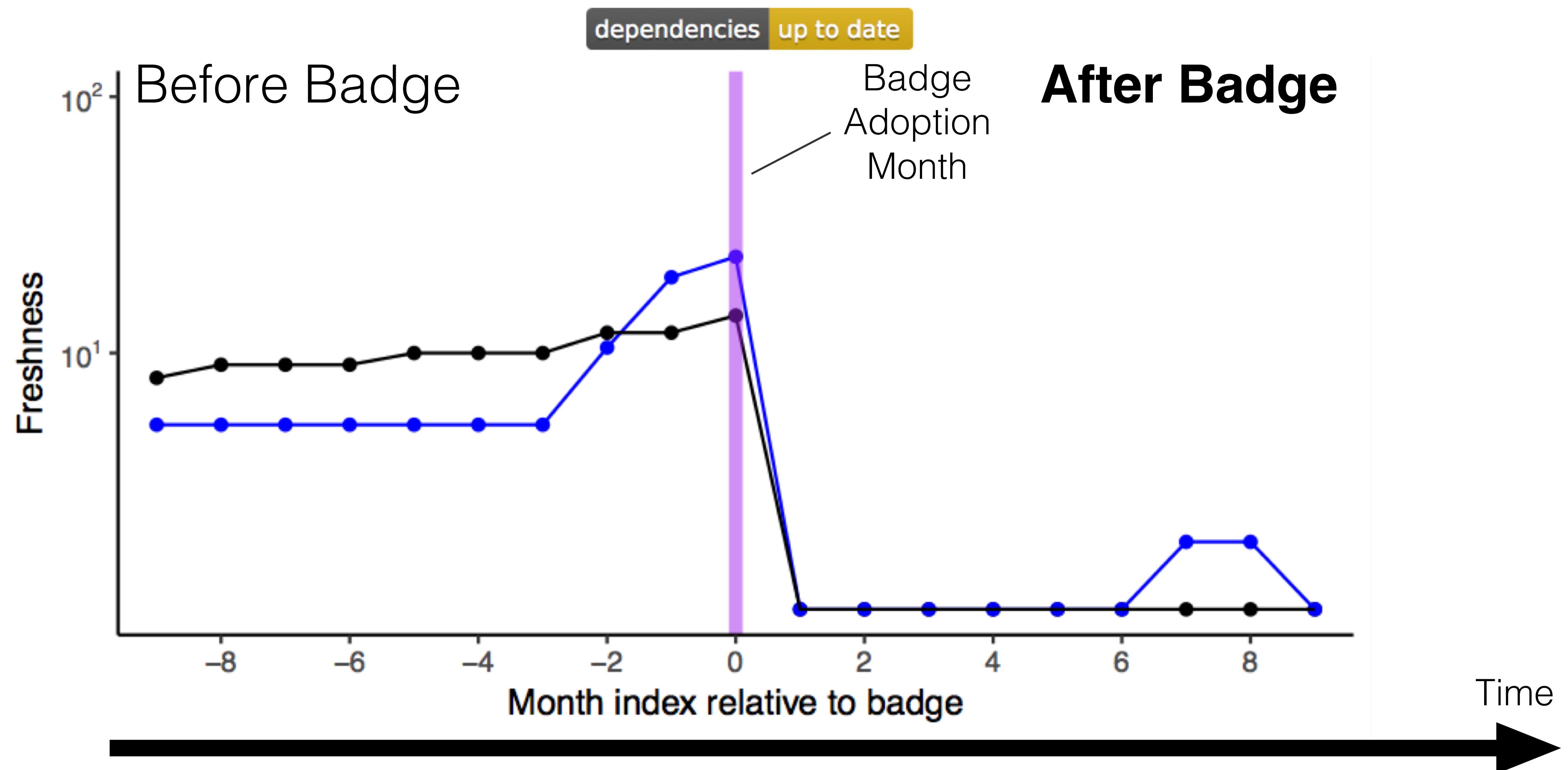


- Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem. Trockman, A., Zhou, S., Kästner, C., and Vasilescu, B. *ICSE 2018*

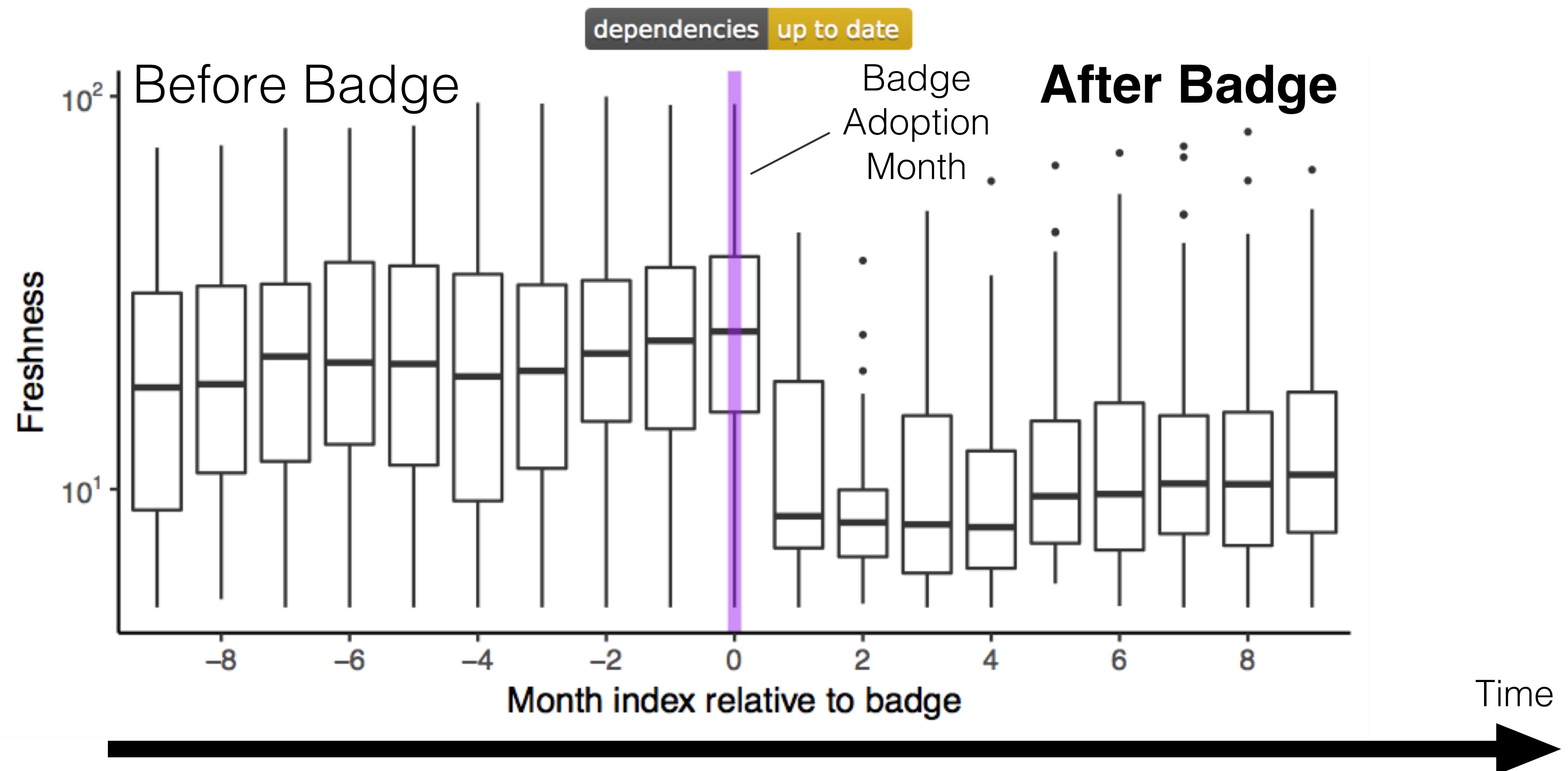
Time Series Analysis



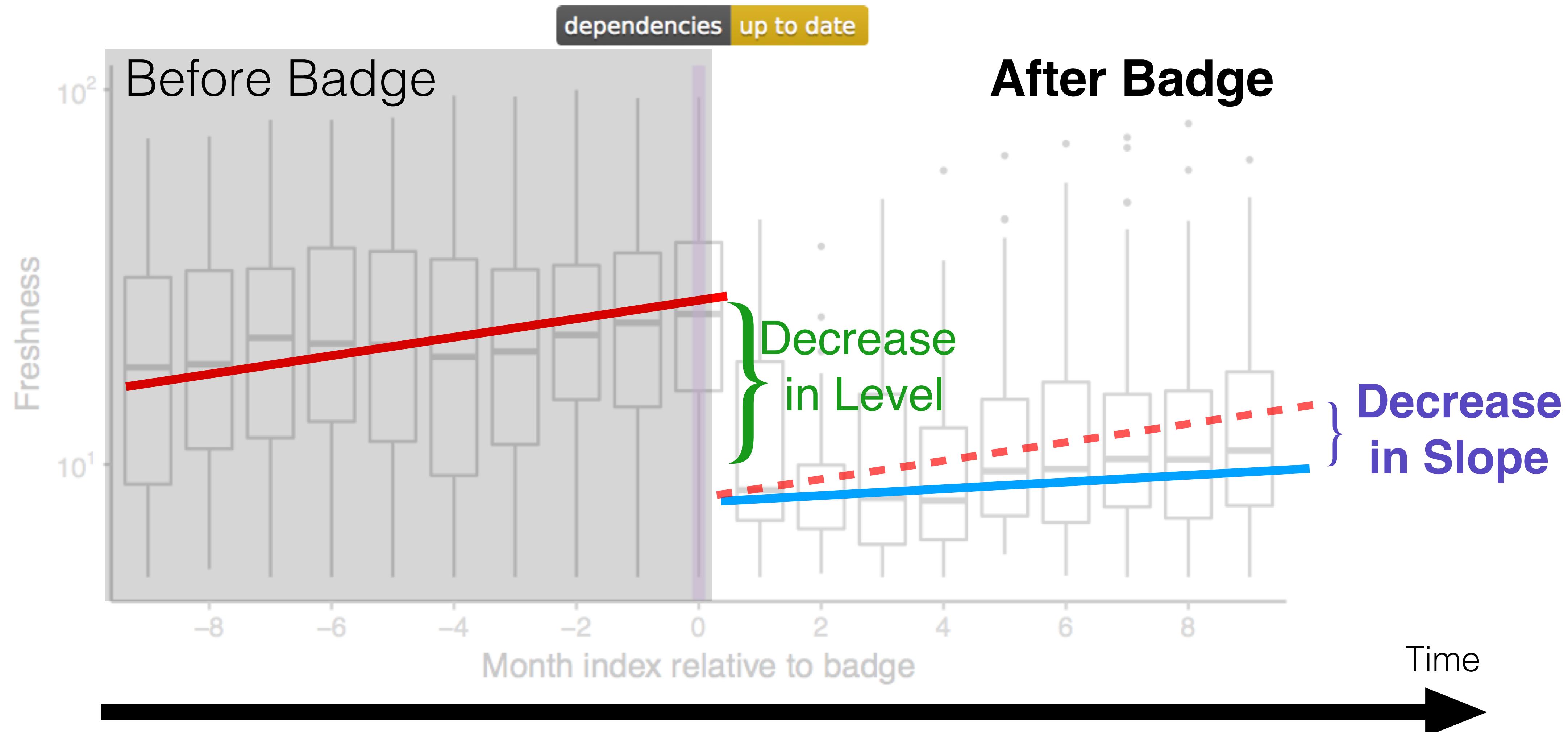
Time Series Analysis



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

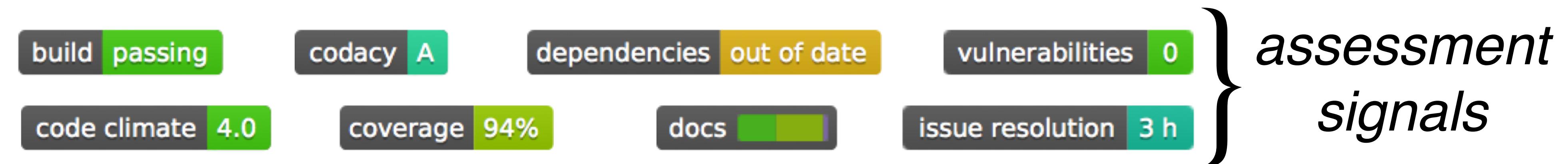
Badges are Reliable Signals

Mostly

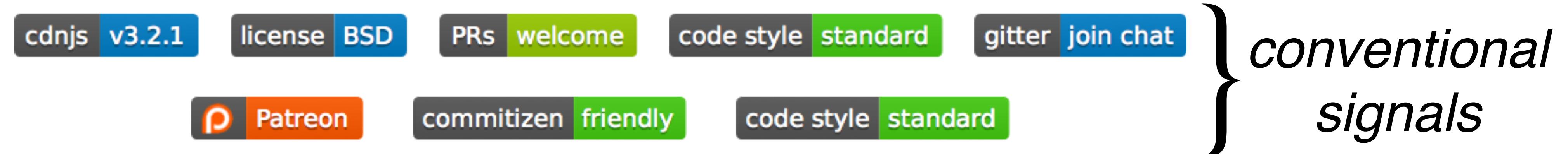


Take-away: Prefer “assessment” badges

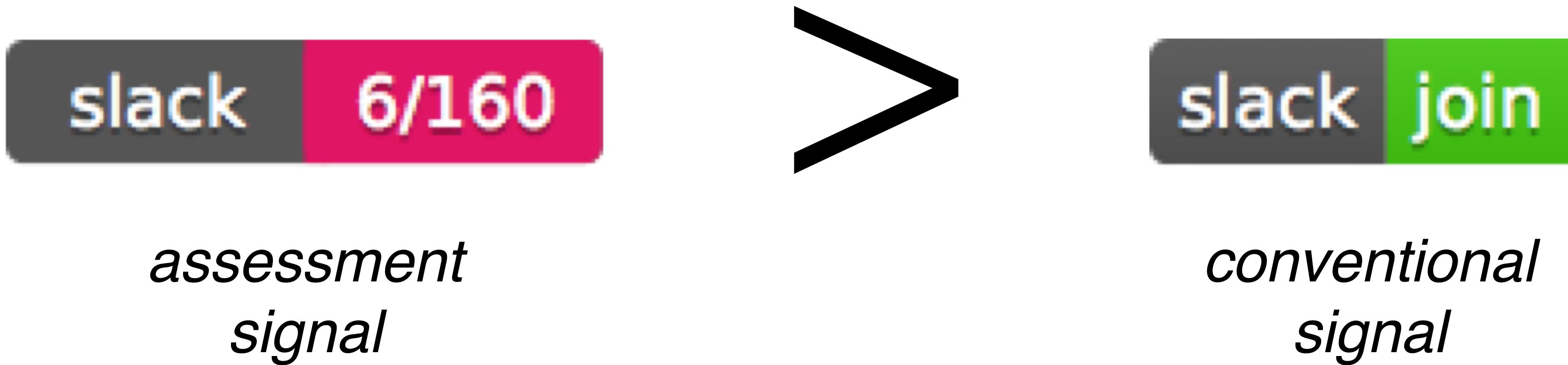
Badges with underlying analyses:



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

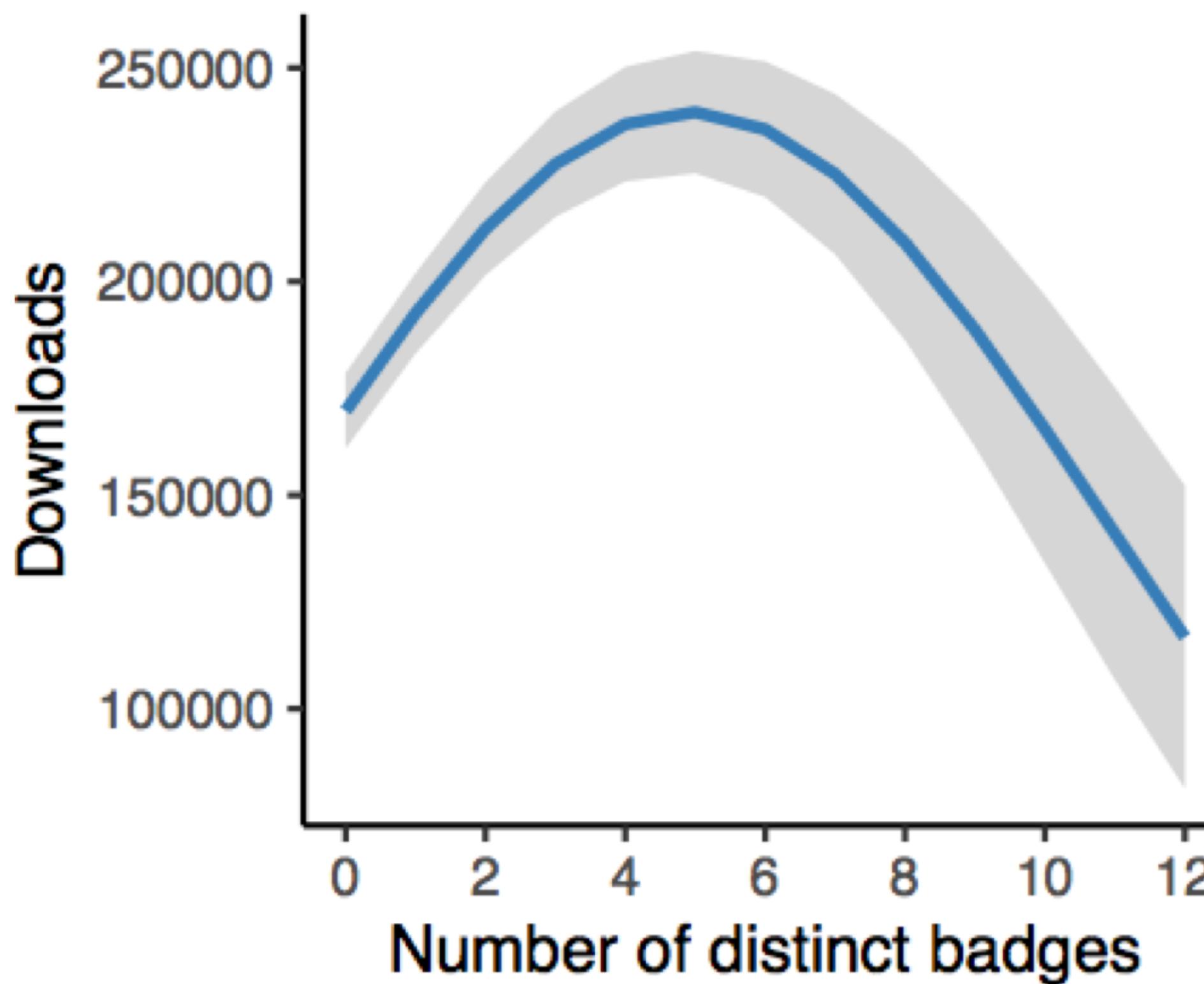


Take-away: Prefer “assessment” badges



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 ?

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*

3. The Dark Side of Transparency

Developers are aware of each other's gender

Survey, 816 responses

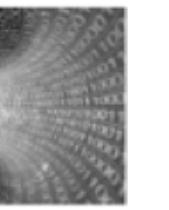
Which of the following characteristics of your team members are you aware of?

- 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”



“Sexist behavior in F/LOSS is as constant as it is extreme”



Article

‘Patches don’t have gender’: What is not open in open source software

new media & society
14(4) 669–683
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461444811422887
nms.sagepub.com


Dawn Nafus
Intel Labs, USA

Abstract
While open source software development promises a fairer, more democratic model of software production often compared to a gift economy, it also is far more male dominated than other forms of software production. The specific ways F/LOSS instantiates notions of openness in everyday practice exacerbates the exclusion of women. ‘Openness’ is a complex construct that affects more than intellectual property arrangements. It weaves together ideas about authorship, agency, and the circumstances under which knowledge and code can and cannot be exchanged. While open source developers believe technology is orthogonal to the social, notions of openness tie the social to the technical by separating persons from one another and relieving them of obligations that might be created in the course of other forms of gift exchange. In doing so, men monopolize code authorship and simultaneously de-legitimize the kinds of social ties necessary to build mechanisms for women’s inclusion.

Pull request acceptance rates are lower when gender is apparent

Article



**'Patches don't have gender':
What is not open in open
source software**

Dawn Nafus
Intel Labs, USA

Abstract
While open source software development promises a fairer, more democratic model of software production often compared to a gift economy, it also is far more male dominated than other forms of software production. The specific ways F/LOSS instantiates notions of openness in everyday practice exacerbates the exclusion of women. 'Openness' is a complex construct that affects more than intellectual property arrangements. It weaves together ideas about authorship, agency, and the circumstances under which knowledge and code can and cannot be exchanged. While open source developers believe technology is orthogonal to the social, notions of openness tie the social to the technical by separating persons from one another and relieving them of obligations that might be created in the course of other forms of gift exchange. In doing so, men monopolize code authorship and simultaneously de-legitimize the kinds of social ties necessary to build mechanisms for women's inclusion.

new media & society
14(4) 669–683
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461444811422887
nms.sagepub.com




Gender differences and bias in open source: pull request acceptance of women versus men

Josh Terrell¹, Andrew Kofink², Justin Middleton², Clarissa Rainear², Emerson Murphy-Hill², Chris Parnin² and Jon Stallings³

¹ Department of Computer Science, California Polytechnic State University—San Luis Obispo, San Luis Obispo, CA, United States

² Department of Computer Science, North Carolina State University, Raleigh, NC, United States

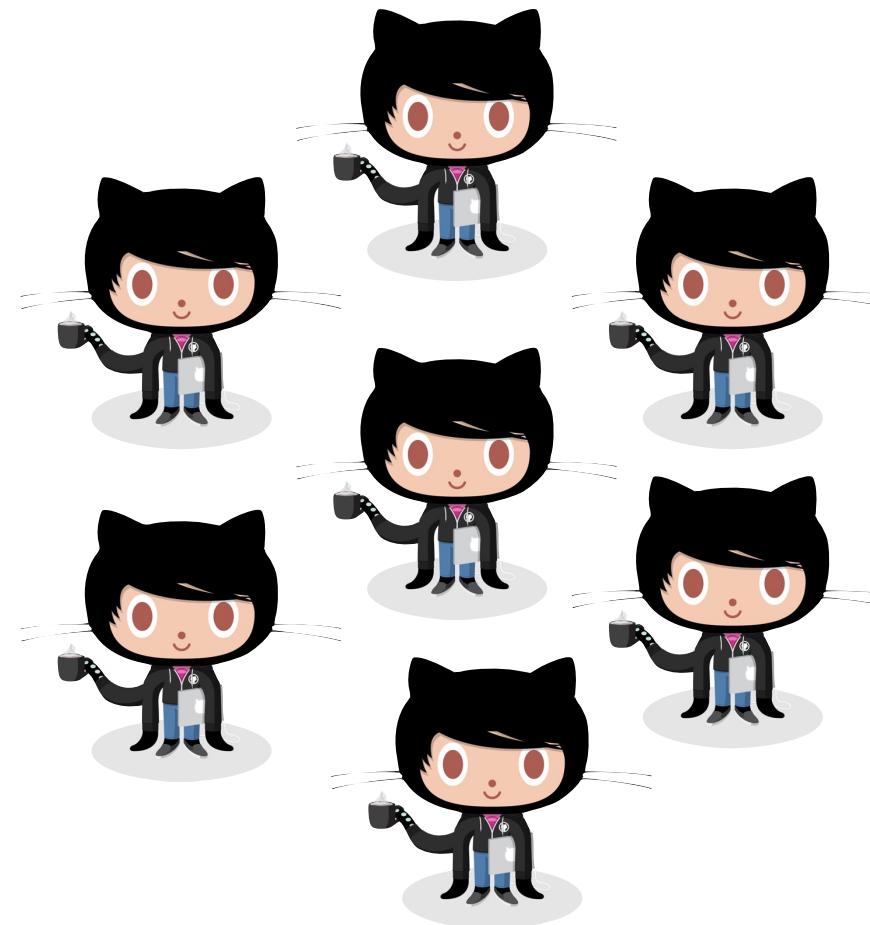
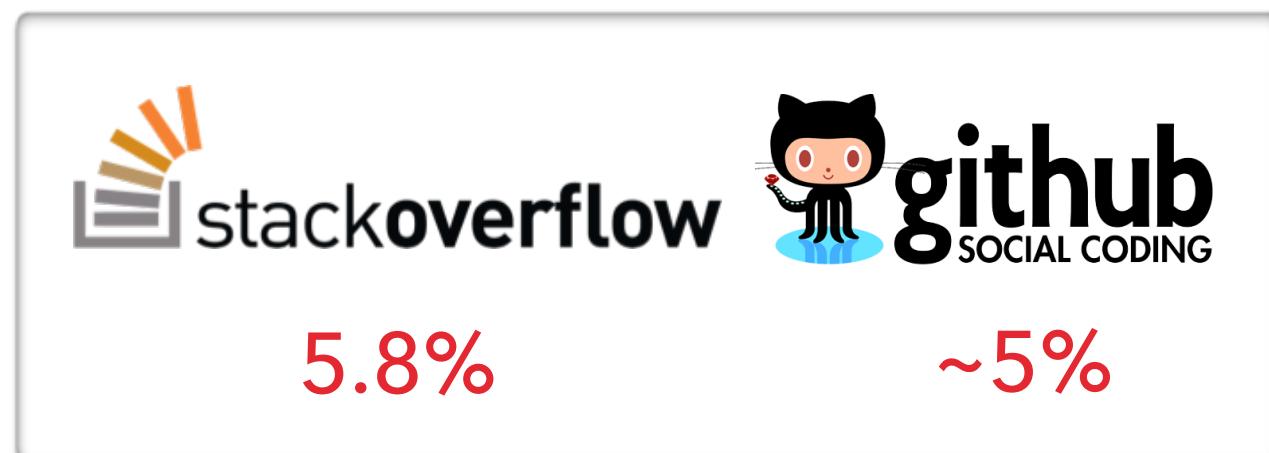
³ Department of Statistics, North Carolina State University, Raleigh, NC, United States

ABSTRACT

Biases against women in the workplace have been documented in a variety of studies. This paper presents a large scale study on gender bias, where we compare acceptance rates of contributions from men versus women in an open source software community. Surprisingly, our results show that women's contributions tend to be accepted more often than men's. However, for contributors who are outsiders to a project and their gender is identifiable, men's acceptance rates are higher. Our results suggest that although women on GitHub may be more competent overall, bias against them exists nonetheless.

Less gender diversity in open source than most places

- Gender representation reality



- Expectation



“More about the contributions to the code than the ‘characteristics’ of the person”

“Any demographic identity is irrelevant”

“Code sees no color or gender”

- FLOSS 2013: A survey dataset about free software contributors: challenges for curating, sharing, and combining G Robles, L Arjona-Reina, B Vasilescu, A Serebrenik, JM Gonzalez-Barahona. *MSR 2014*
- Google Diversity (2015) www.google.com/diversity/index.html#chart
- Inside Microsoft (2015) <https://goo.gl/nT4YiI>

- Exploring the data on gender and GitHub repo ownership Alyssa Frazee. <http://alyssafrazee.com/gender-and-github-code.html>
- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries) <http://stackoverflow.com/research/developer-survey-2015#profile-gender>

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

Again, lots of anecdotes

Experiences working in a diverse team

“code sees no color or gender”

Meritocracy; no effects of diversity

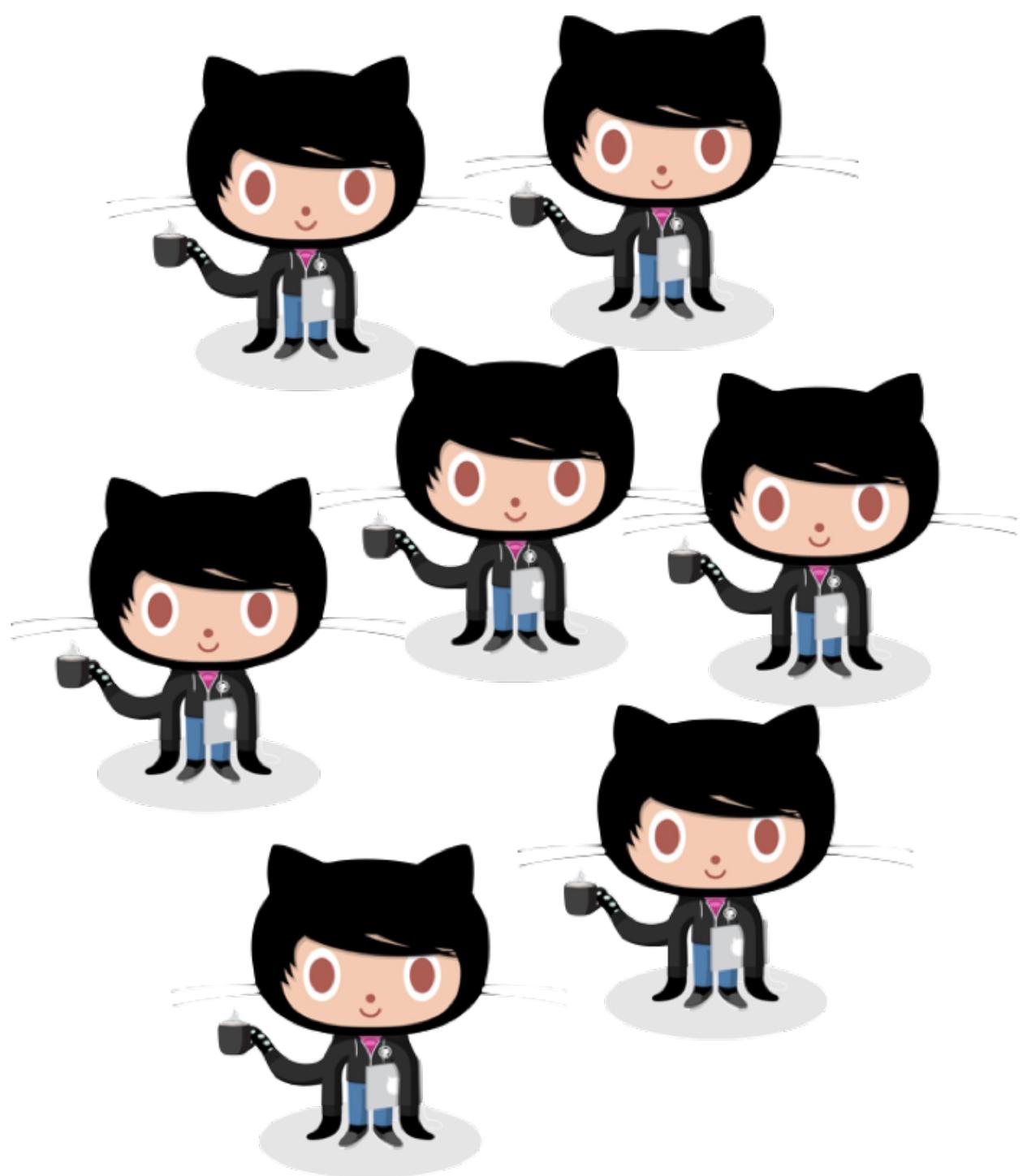
“diverse viewpoints often lead to **lively discussions and new ideas**”

Positive effects of diversity

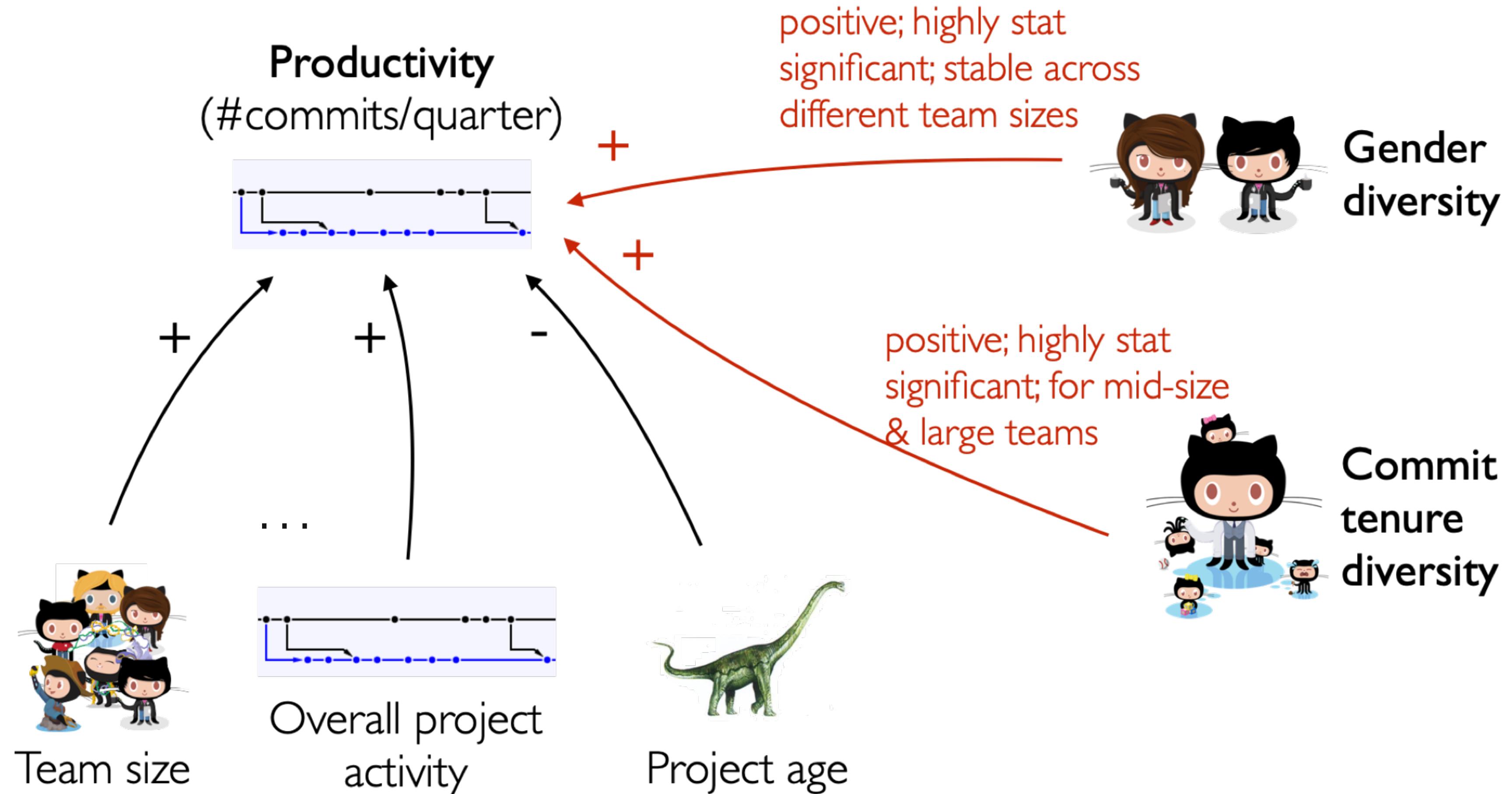
“I have used a **fake GitHub handle** (my normal GitHub handle is my first name, which is a distinctly female name) **so that people would assume I was male**”

Negative effects of diversity

Which tends to be more effective, on average?



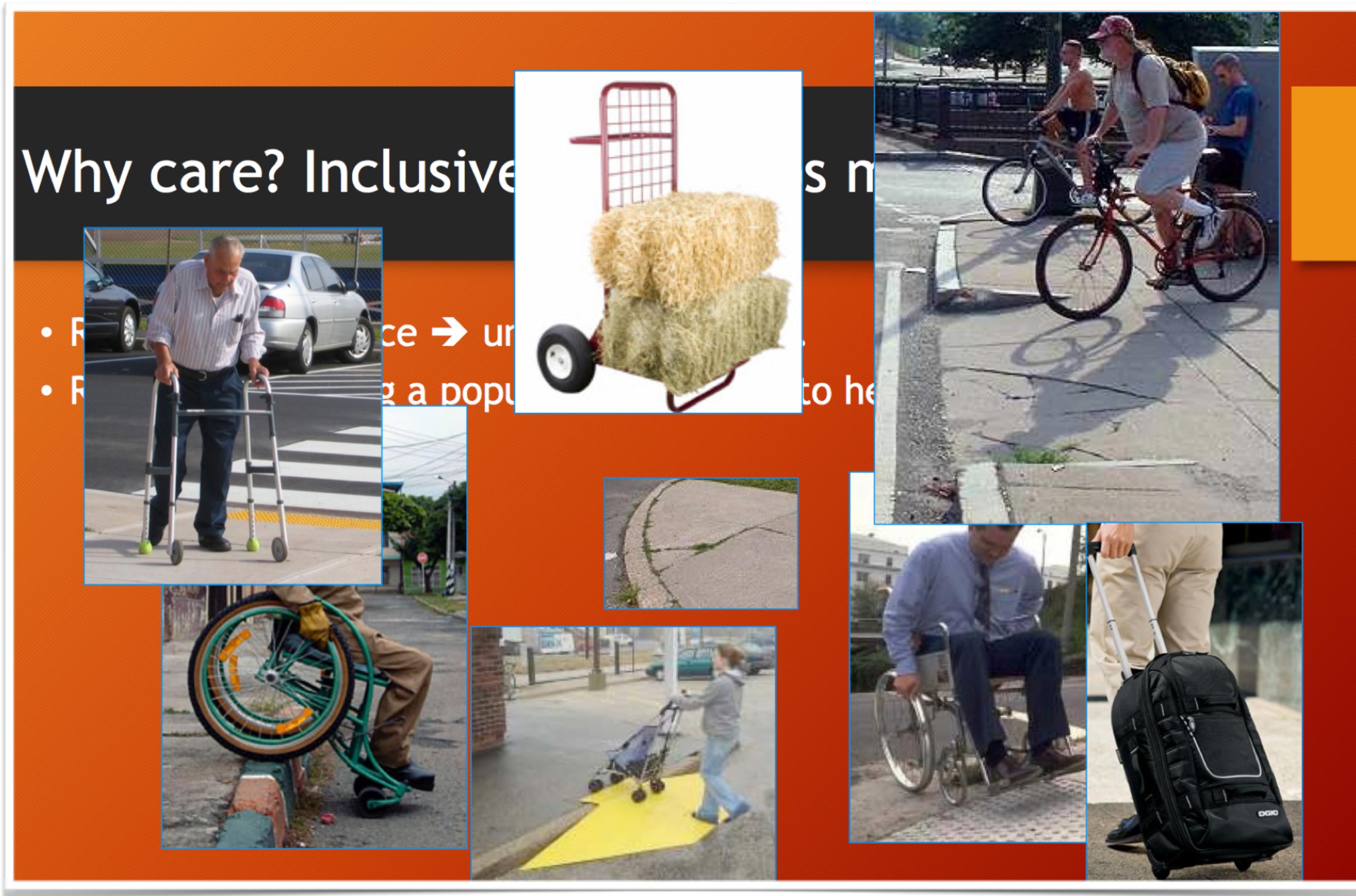
Increased diversity correlates to higher productivity



- 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

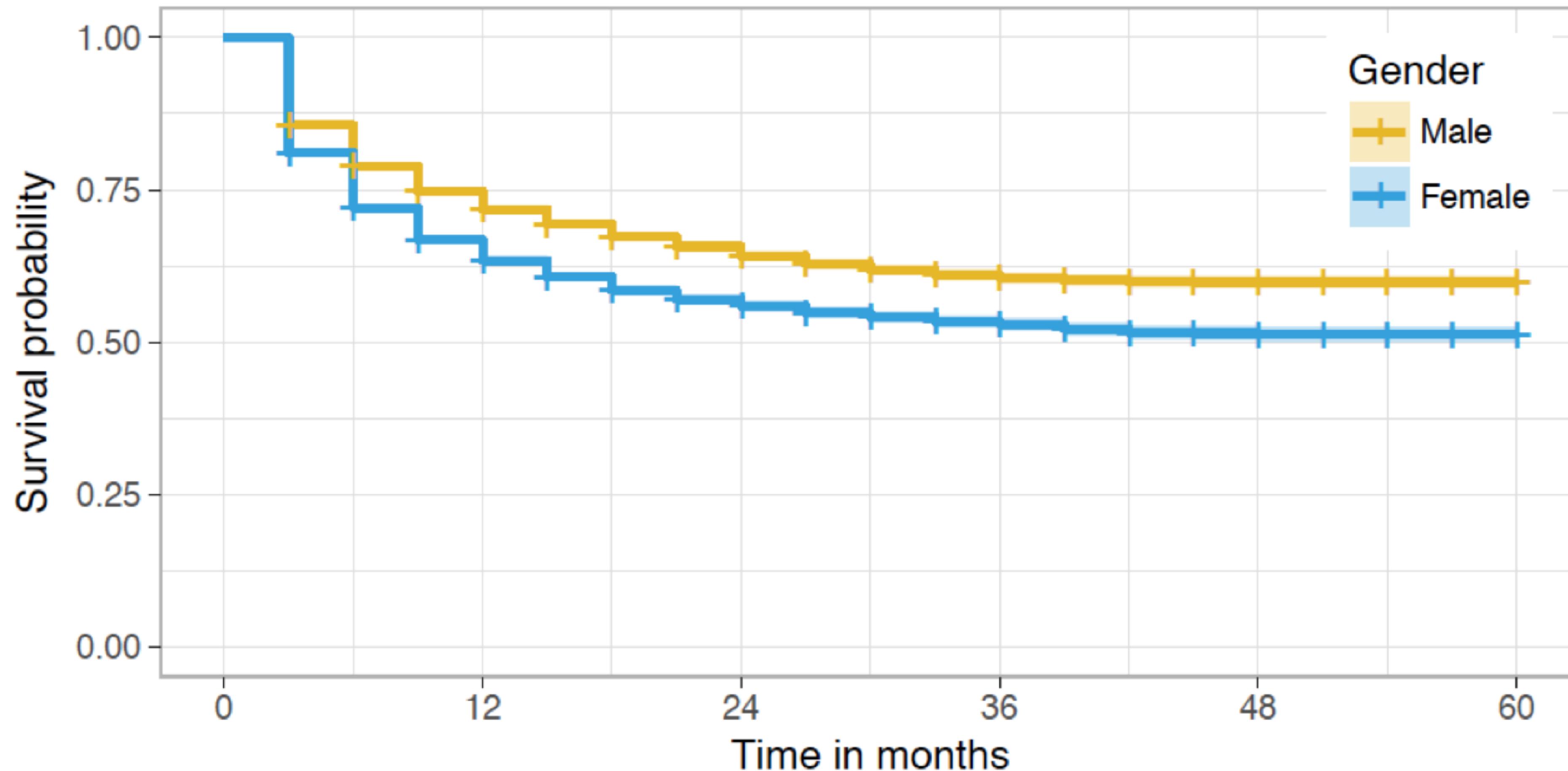


© Anita Sarma & Margaret Burnett, Oregon State U

4.

Dropout and retention

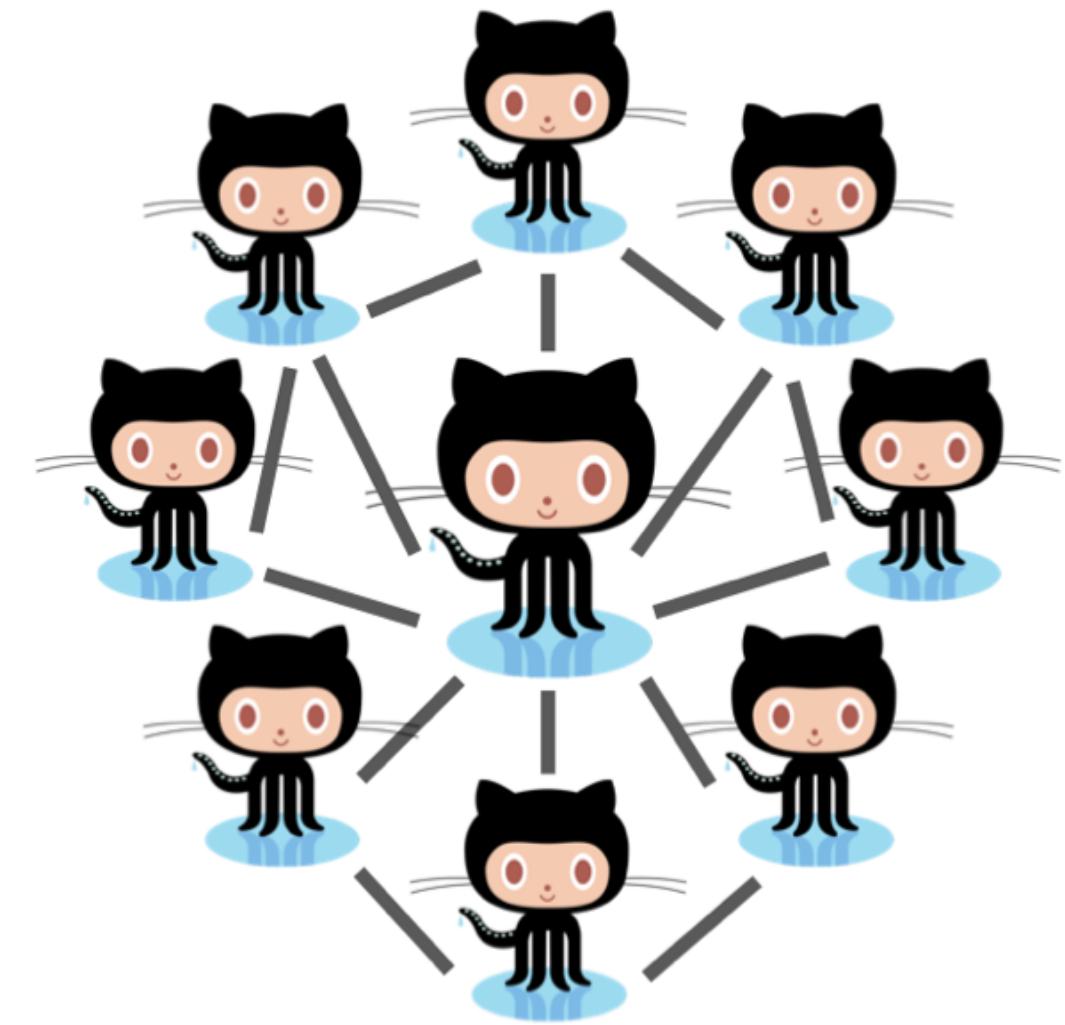
Women on GitHub disengage earlier than men



- Going Farther Together: The Impact of Social Capital on Sustained Participation in Open Source. Qiu, H.S., Nolte, A., Brown, A., Serebrenik, A., and Vasilescu, B. *ICSE 2019*

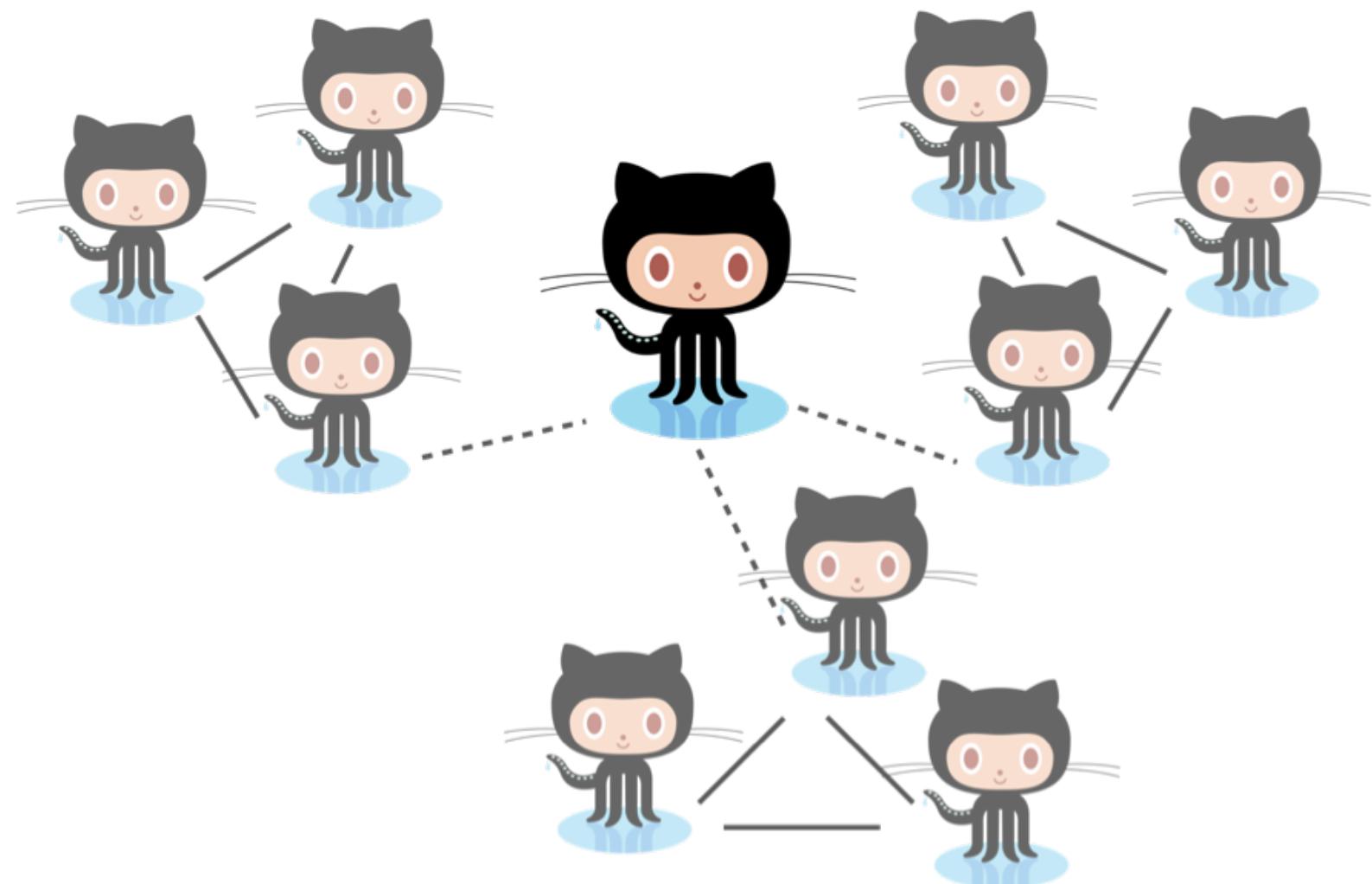
Social capital theory explains long-term engagement

Bonding social capital:
benefiting from strongly
connected network



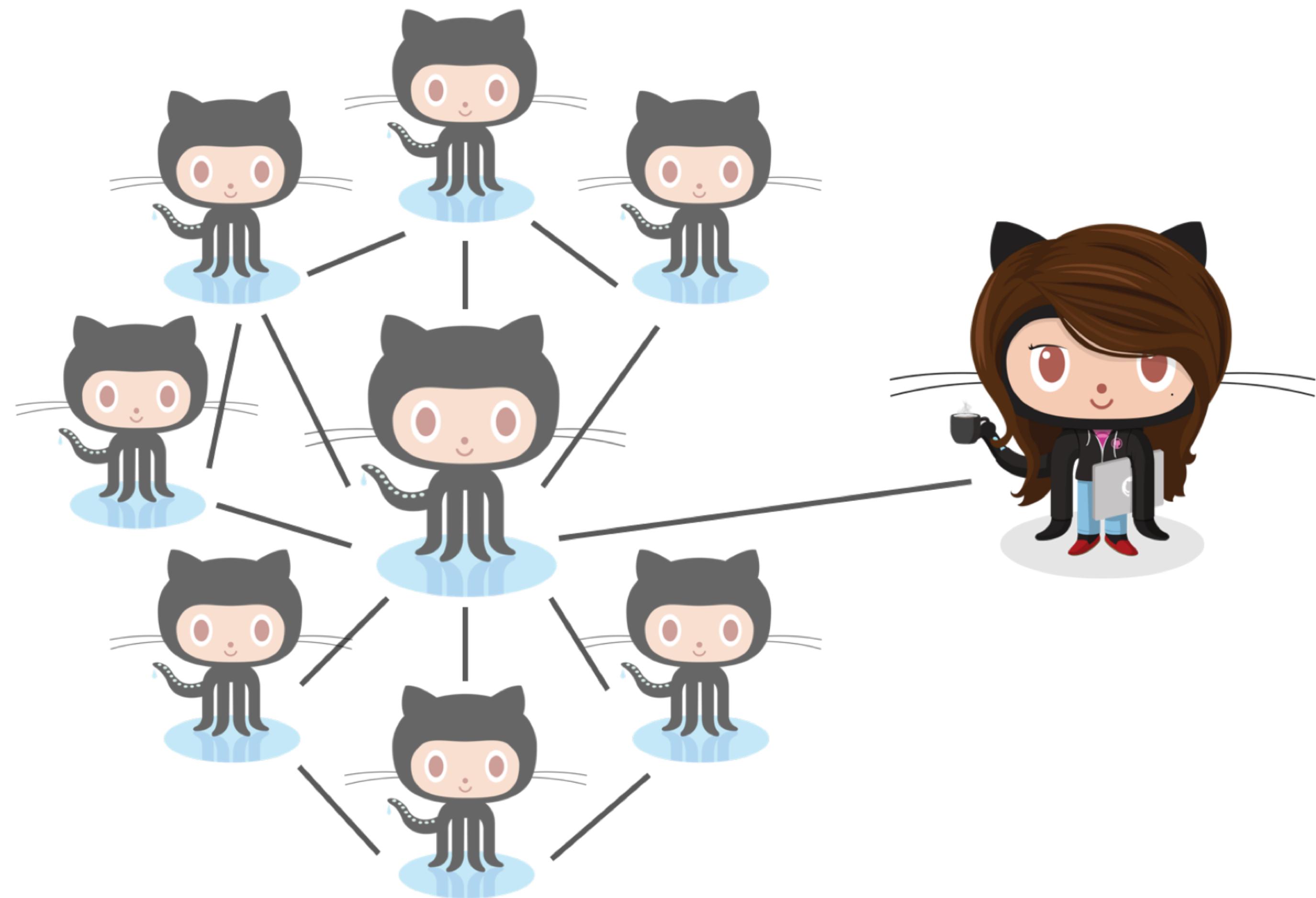
Willingness to continue
(Coleman, 1990)

Bridging social capital:
benefiting from network
with diverse info



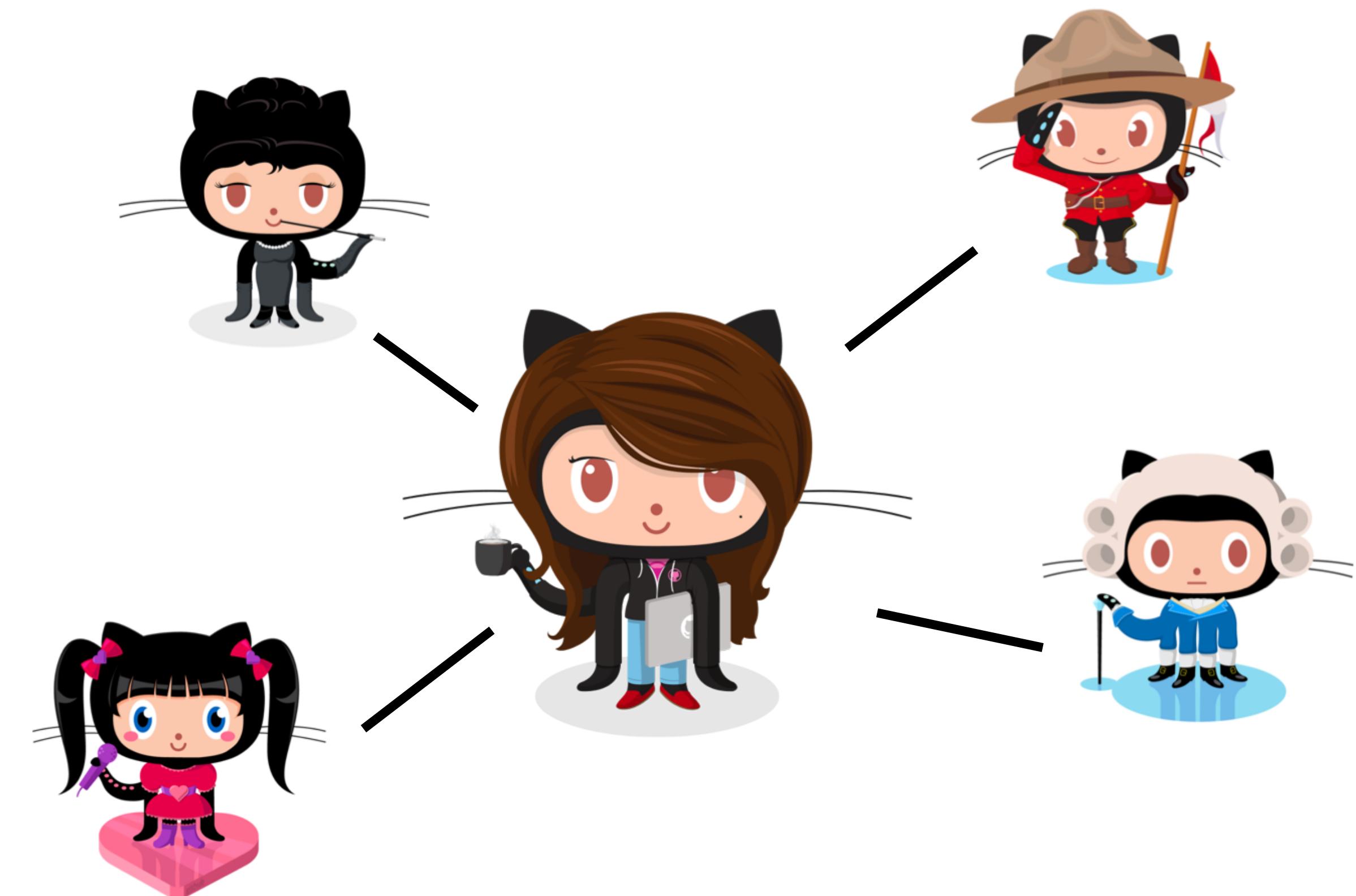
Opportunity to continue
(Burt, 1998, 2001)

Cohesive networks might foster discrimination / exclusion

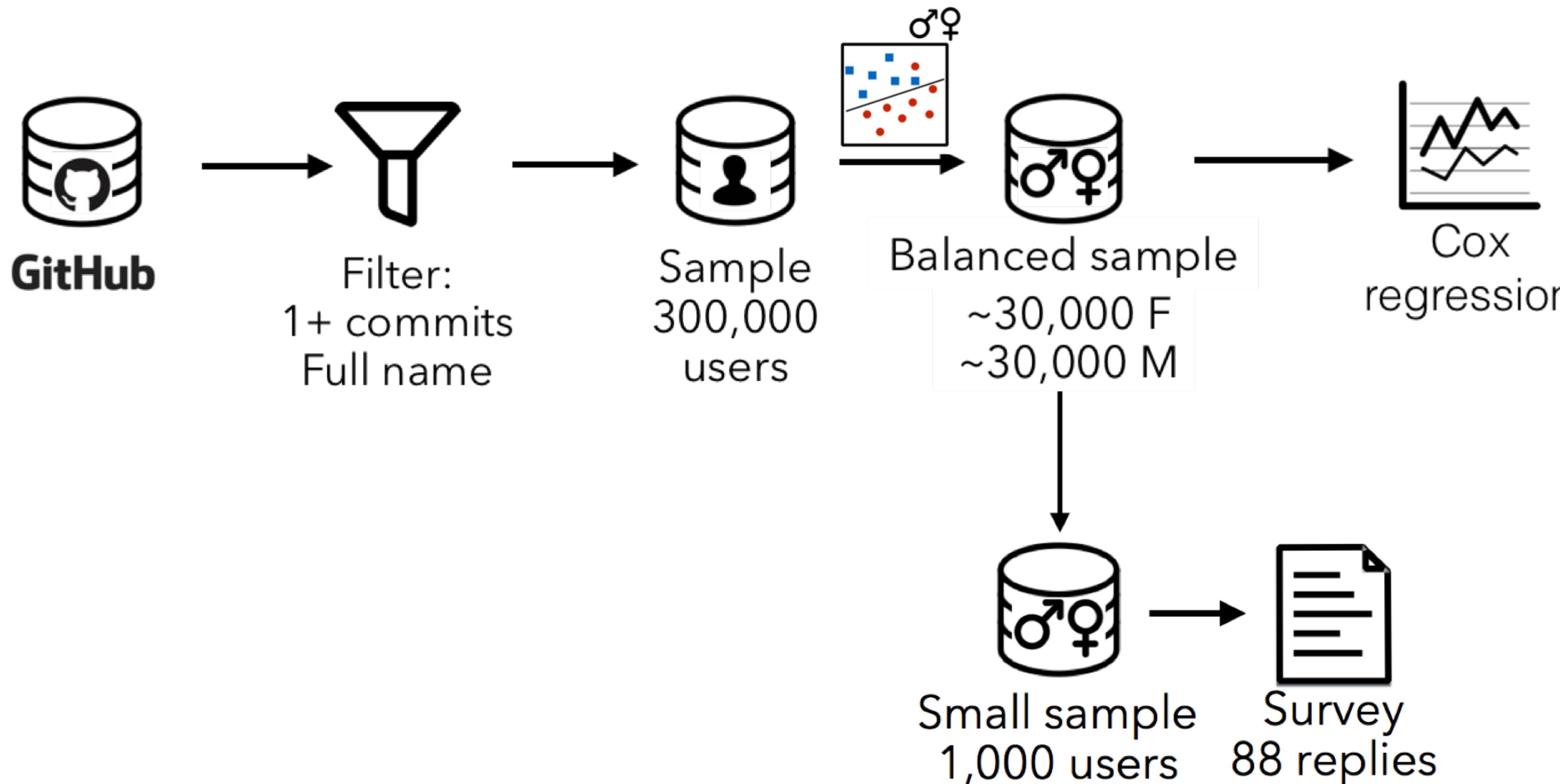


Being part of teams with more diverse information ~ more prolonged engagement, esp. for women

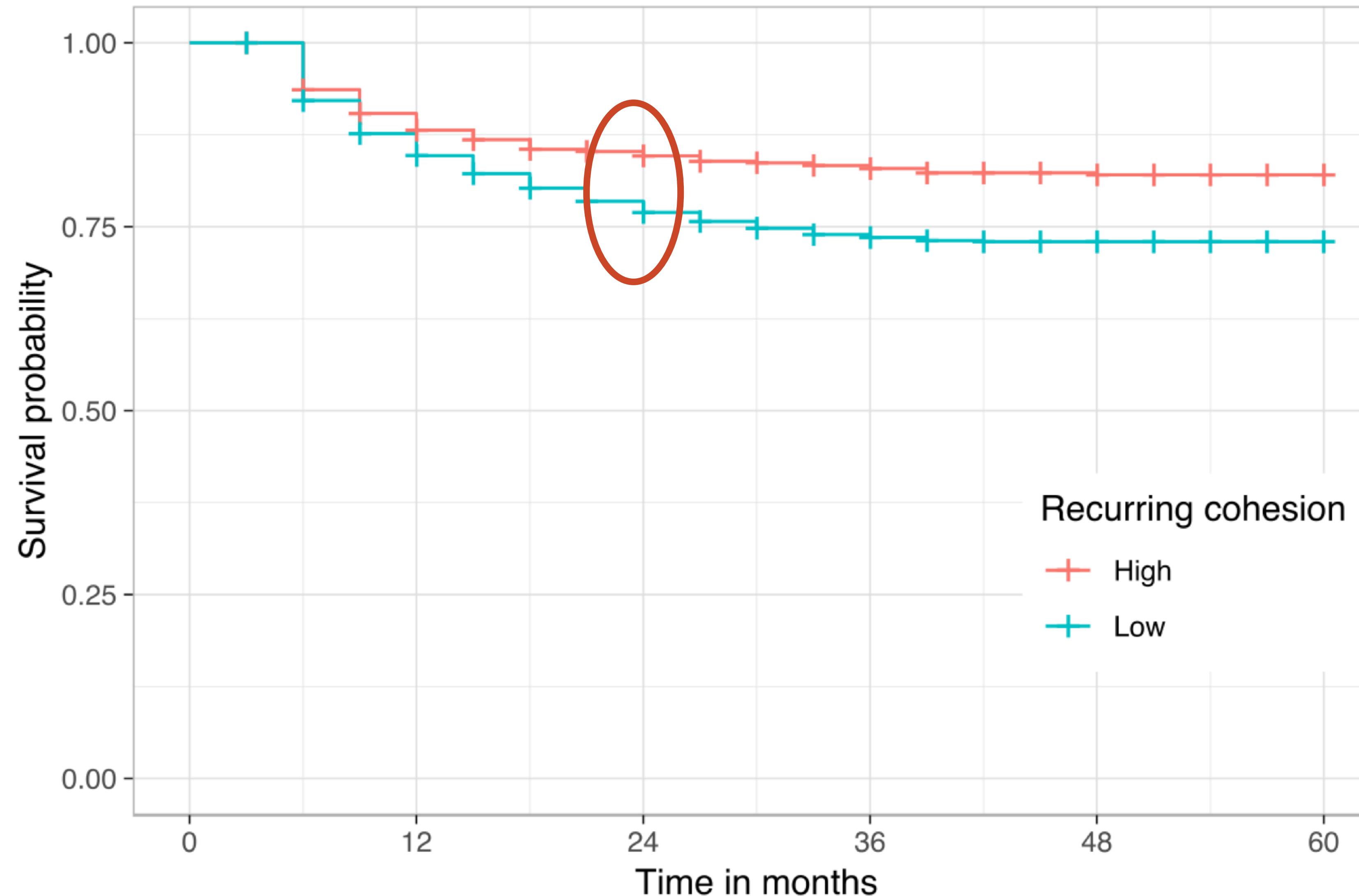
Information diversity should
reduce the risk of demographic-
based echo chambers.



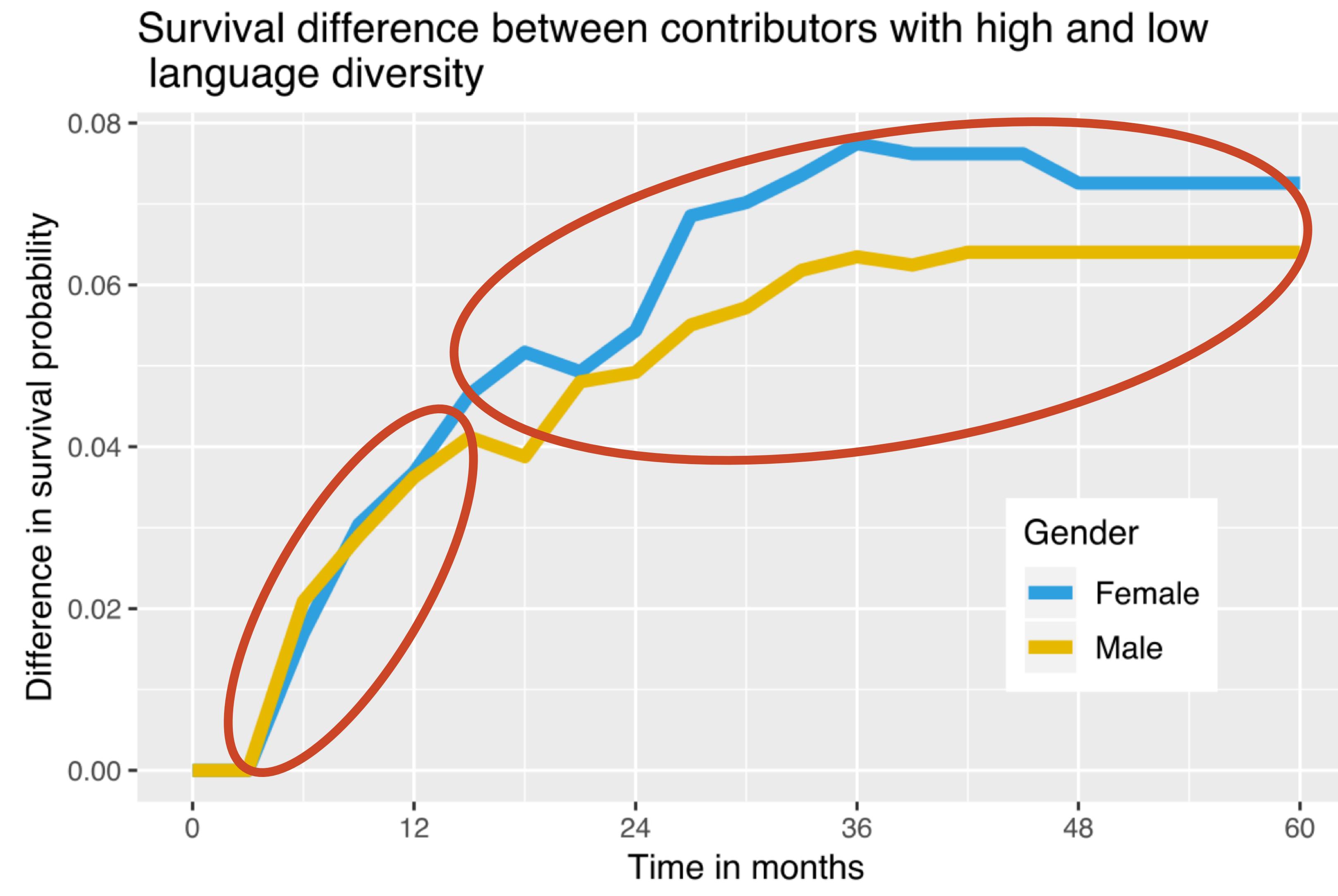
Large-scale mixed-methods study



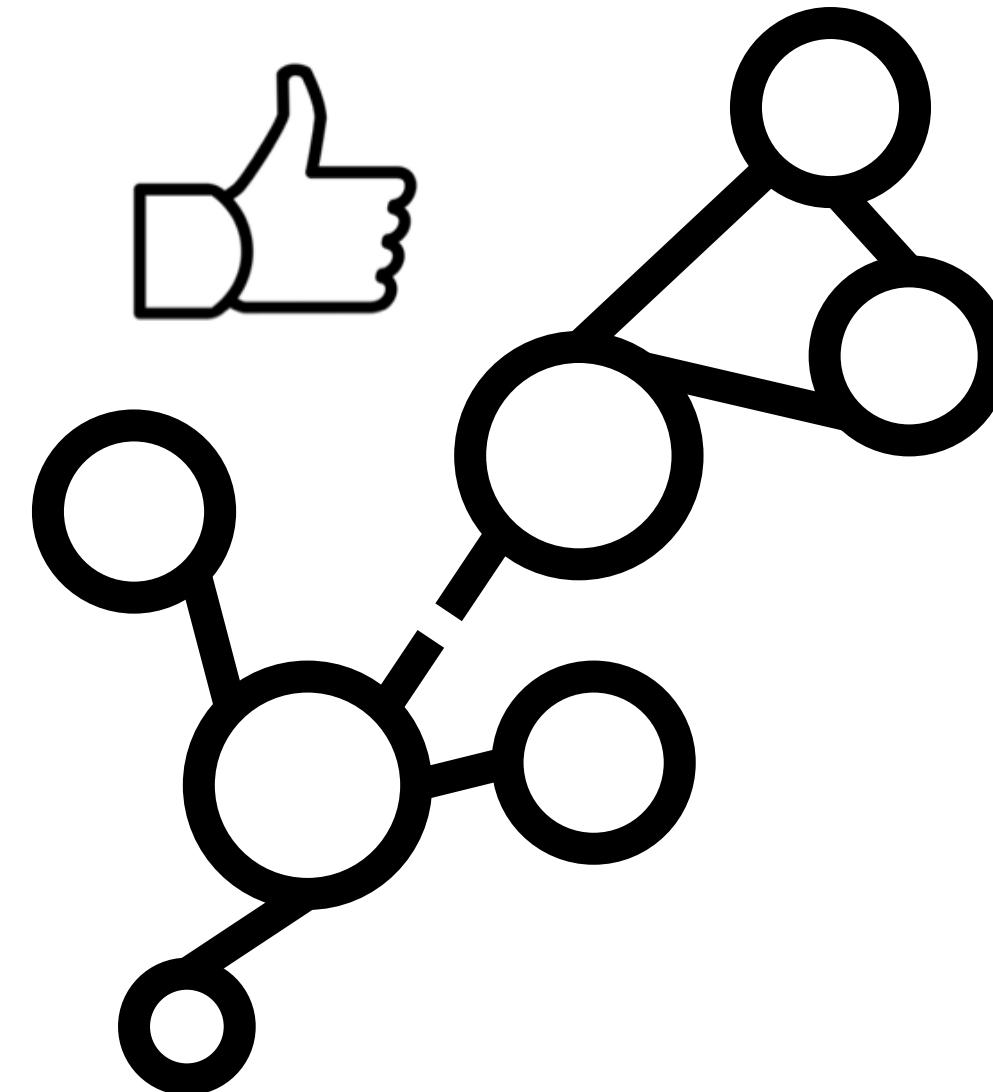
More social capital ~ more prolonged engagement



Women in language- (informationally-) diverse teams disengage at lower rates



Take away: Invest in building social capital & Foster informationally diverse teams



Recommend projects that can help build social capital



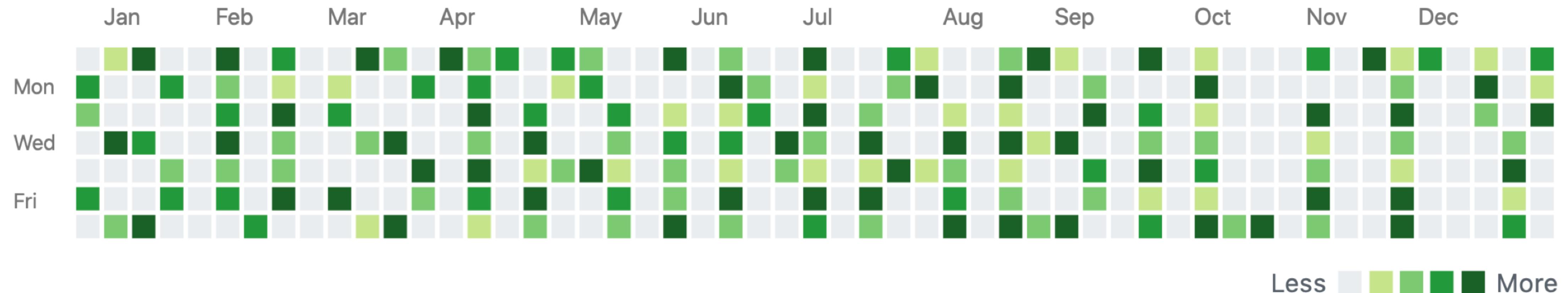
mentorship 10 mentors

Find relevant mentorship



Signal social capital moderators

Summary



Less More

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.



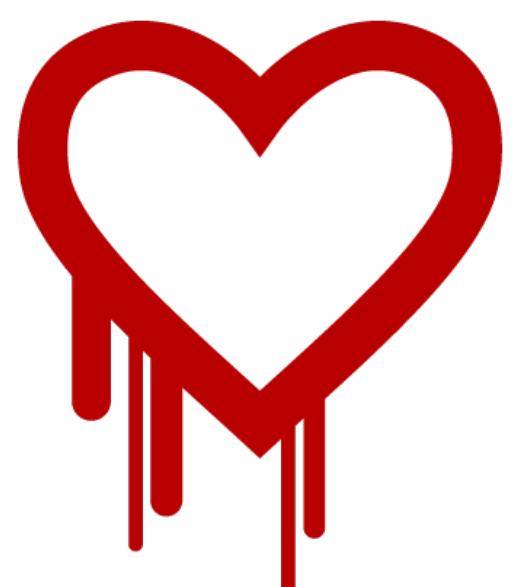
By Zack Whittaker | September 14, 2017 -- 01:27 GMT (18:27 PDT) | Topic: Security



The Heartbleed Bug

The Heartbleed Bug is a serious vulnerability in the popular OpenSSL cryptographic software library. This weakness allows stealing the information protected, under normal conditions, by the SSL/TLS encryption used to secure the Internet. SSL/TLS provides communication security and privacy over the Internet for applications such as web, email, instant messaging (IM) and some virtual private networks (VPNs).

The Heartbleed bug allows anyone on the Internet to read the memory of the systems protected by the vulnerable versions of the OpenSSL software. This compromises the secret keys used to identify the service providers and to encrypt the traffic, the names and passwords of the users and the actual content. This allows attackers to eavesdrop on communications, steal data directly from the services and users and to impersonate services and users.



What leaks in practice?

We have tested some of our own services from attacker's perspective. We attacked ourselves from outside, without leaving a trace. Without using any privileged information or credentials we were able to steal from ourselves the secret keys used for our X.509 certificates, user names and passwords, instant messages, emails and business critical documents and communication.

How to stop the leak?

As long as the vulnerable version of OpenSSL is in use it can be abused. Fixed OpenSSL (<https://www.openssl.org/news/secadv/20140407.txt>) has been released and now it has to be deployed. Operating system vendors and distribution, appliance vendors, independent software vendors have to adopt the fix and notify their users. Service providers and users have to install the fix as it becomes available for the operating systems, networked appliances and

We have seen...

- Limitations of donations as a sustainable funding source
- Badges as a transparent signaling mechanism
- A dark side to transparency
- A bit of empirical evidence in favor of diversity
- Social capital theory suggesting a path to improve retention

We have seen...

- Analysis of terabytes of public trace data
- Mixed methods research
- The slow process from anecdotal evidence to evidence-based recommendations
- Eventual goal: intentional design of tools, communities, and interventions

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)

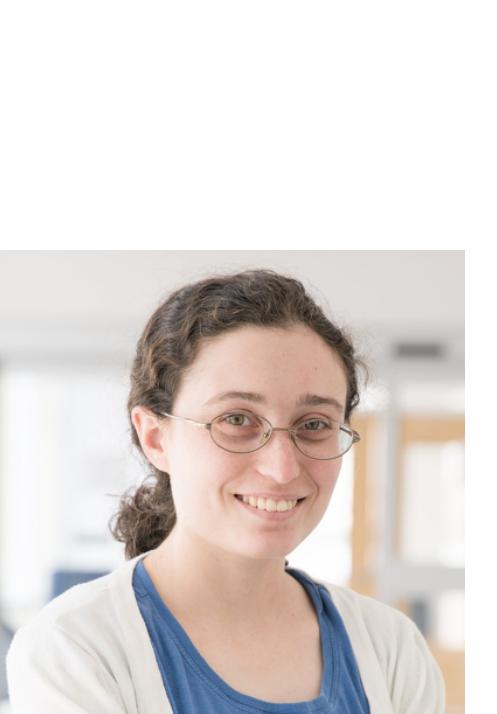
Diversity and inclusion

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

Acknowledgements



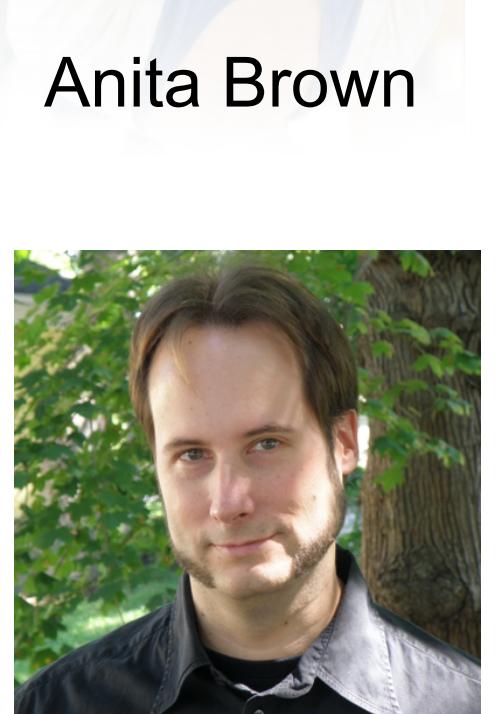
Courtney Miller



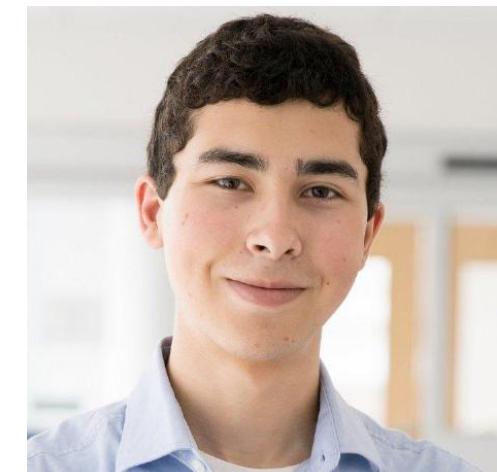
Anita Brown



Audris Mockus



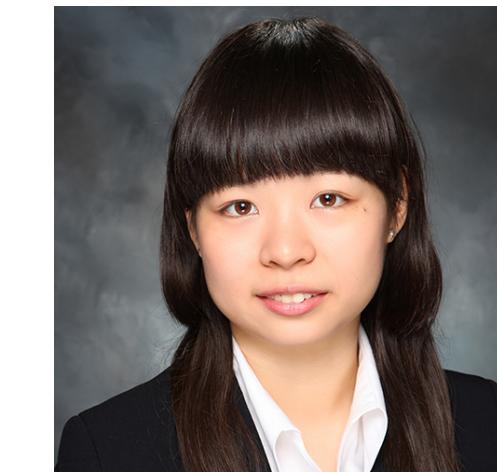
Alex Nolte



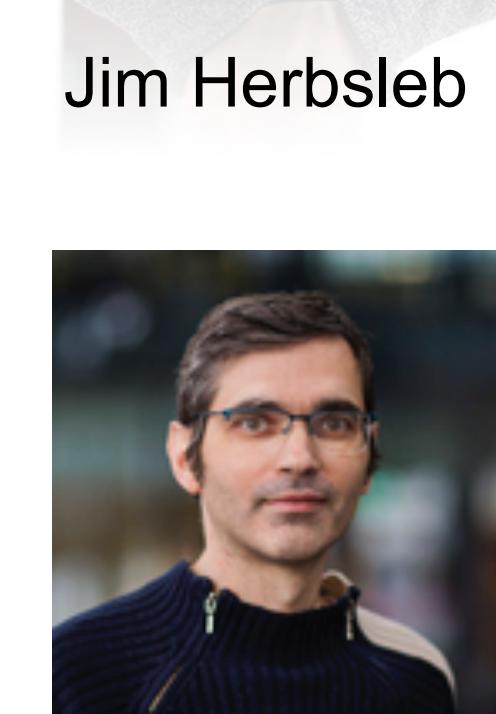
Asher Trockman



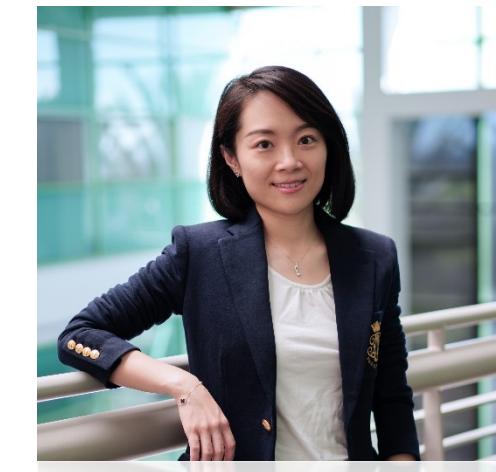
Jim Herbsleb



Sophie Qiu



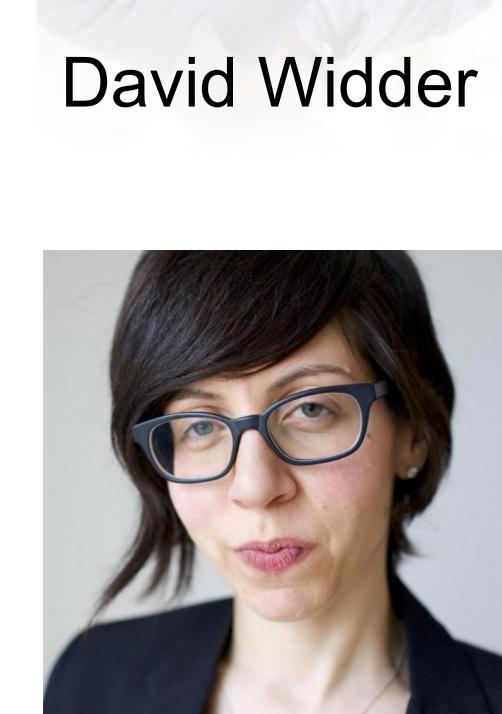
Alex Serebrenik



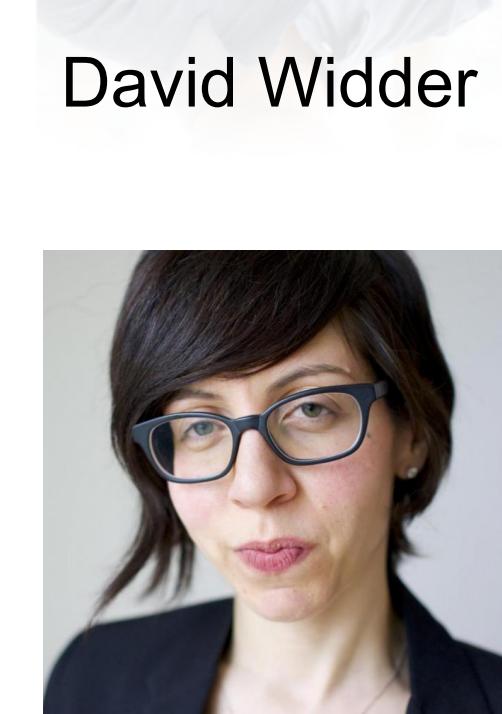
Shurui Zhou



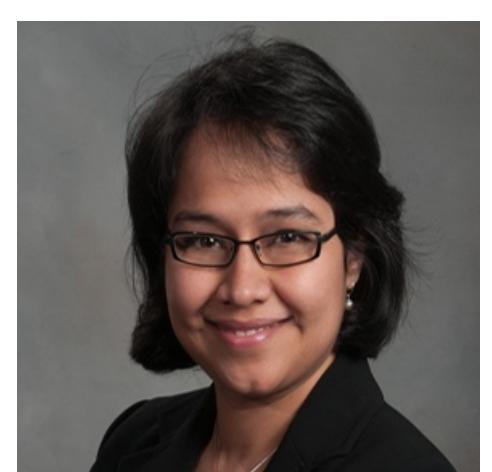
David Widder



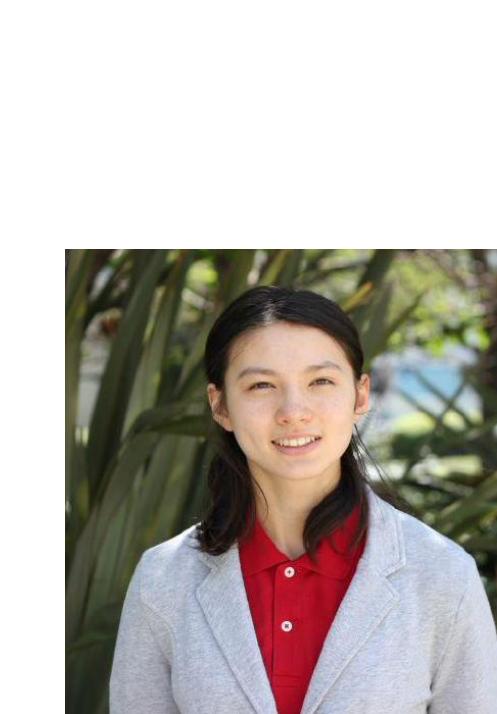
Marat Valiev



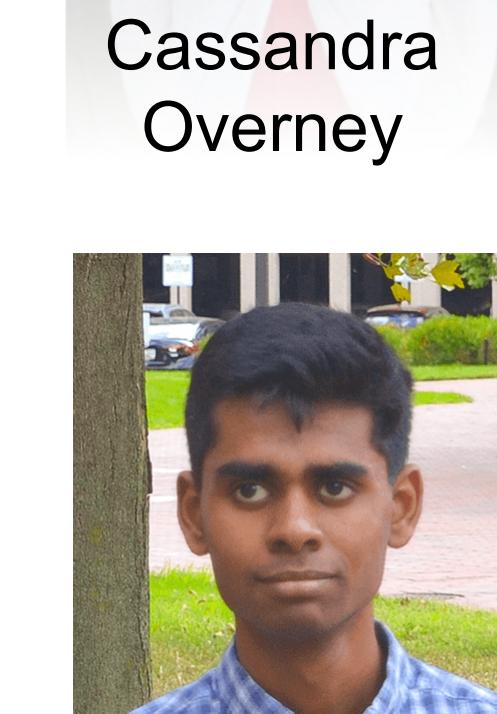
Laura Dabbish



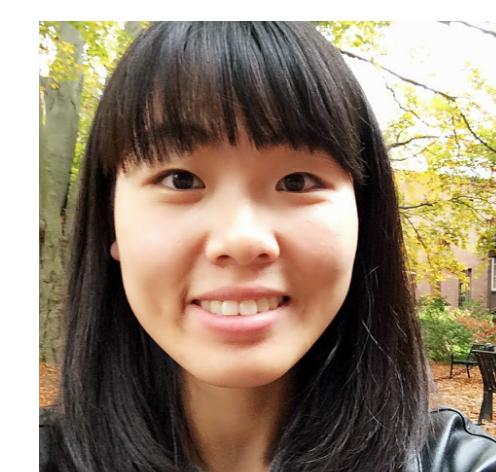
Anita Sarma



Cassandra Overney



Naveen Raman



Michelle Cao



Christian
Kästner



National Science
Foundation

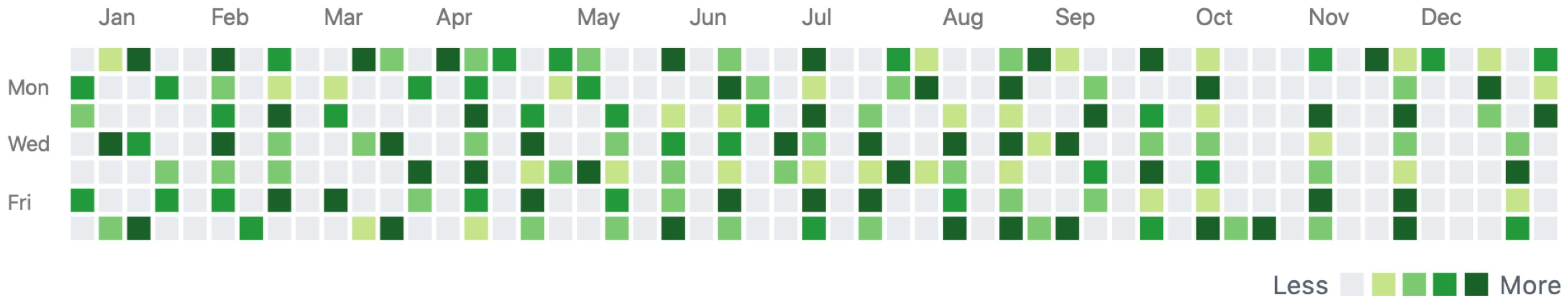


Alfred P. Sloan
FOUNDATION



FORDFOUNDATION

What can analyzing tens of terabytes of public trace data tell us about open source



Bogdan Vasilescu
@b_vasilescu