



Measuring Open-Source Software as an Intangible, Digital Asset using GitHub

Sixth World KLEMS Conference

Digital Economy Session March 16, 2021

Carol Robbins, NCSES

National Center for Science and Engineering Statistics
Social, Behavioral and Economic Sciences
National Science Foundation

Collaborators

Gizem Korkmaz	Associate Professor, Biocomplexity Institute, UVA
---------------	--

Ledia Guci	Senior Analyst, NCSES, NSF
------------	----------------------------

Bayoán Santiago Calderón	Postdoctoral Research Associate, Biocomplexity Institute, UVA
-----------------------------	--

Brandon Kramer	Postdoctoral Research Associate, Biocomplexity Institute, UVA
----------------	--

Disclaimer: The views expressed in this paper are those of the authors and not necessarily those of their respective institutions.

Acknowledgments: This material is based on work supported by U.S. Department of Agriculture (58-3AEU-7-0074) and the National Science Foundation (Contract #49100420C0015)

Open-Source Software: an Intangible Digital Asset

“Open Source Software (OSS) is a computer software, with its source code made available with a license, in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose.” ([Open Source Initiative](#))

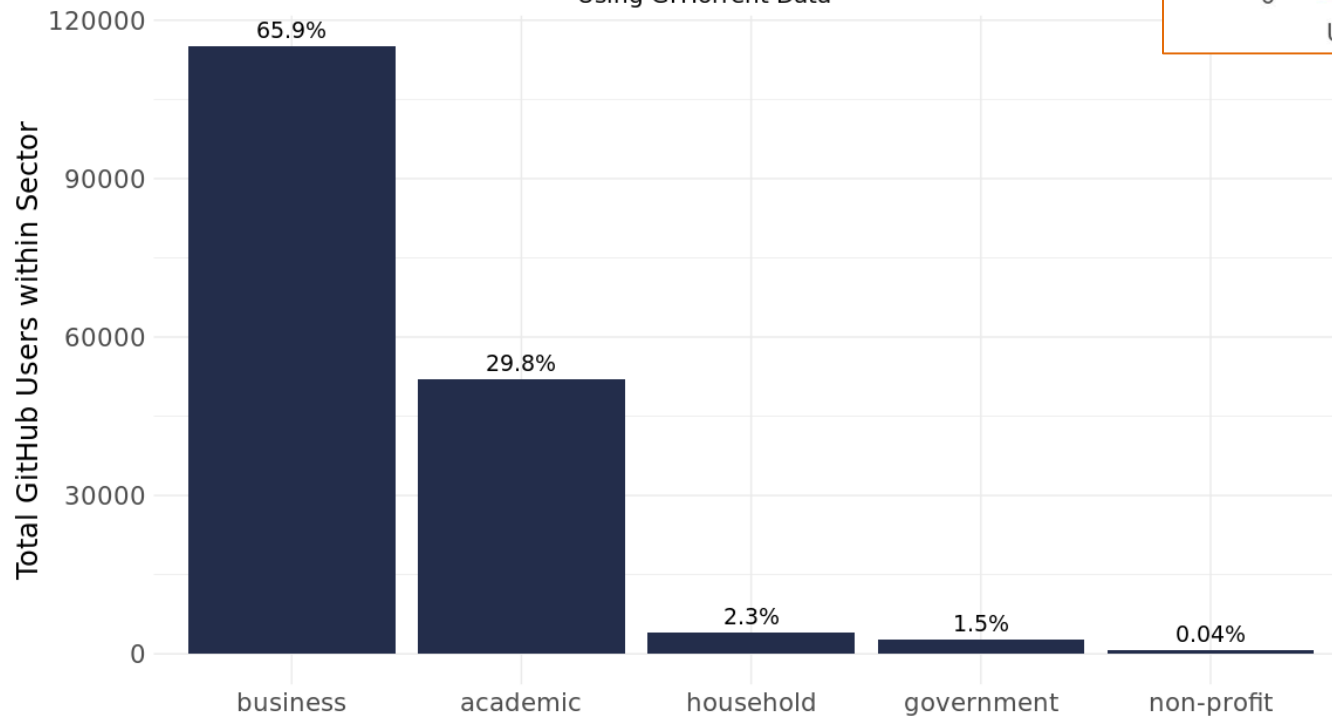
Developed, maintained and extended by:

- *universities* (e.g., Stanford, MIT, UC, Berkeley)
- *businesses* (e.g., Microsoft, Google)
- *government research institutions* (e.g., Sandia National Lab)
- *Nonprofits*
- *individuals*





GitHub Users Classified by Economic Sector
Using GHTorrent Data



Where is it coming from and who is creating it?

Overview

- Motivation
- Knowledge outputs and the System of National Accounts
- Data Discovery
- Quantity/Volume
- Sector and Country
- Where we are headed: time series investment and capital stock

NCSES Data on Human Capital, R&D, and Innovation

NCSES's mandate is the collection, interpretation, analysis, and dissemination of objective data on the science and engineering enterprise.

NCSES's mission:

- Research and Development
- The science and engineering workforce
- U.S. competitiveness in science, engineering, technology, and R&D
- The condition and progress of STEM education in the United States

Data Products include:

- Workforce Statistics
- R&D Statistics
- Business Innovation Statistics
- Indicators of Research, Invention, and Innovation

2018 Oslo Manual Promotes Bringing Innovative Knowledge into the SNA

- Integrating Innovation Data with SNA sources
- 2018 Revision of Oslo Manual
 - SNA framework recommended for collection of innovation statistics
 - Use SNA terminology where applicable
 - Innovation in all SNA sectors should follow SNA
 - Business
 - General government
 - Non-profit institutions serving households
 - Household
 - Not going to happen all at once
 - Universities

CHAPTER 2. CONCEPTS FOR
MEASURING INNOVATION | 55
OSLO MANUAL 2018 ©
OECD/EUROPEAN UNION 2018



TAMAYO, Rufino (1899-1991)
PROMETHEUS BRINGING FIRE TO MANKIND, 1958
Fresco, signed lower right 'Tamayo 9-58', 500 x 450 cm
<http://www.unesco.org/artcollection/>

Inspiration

- Corrado, Hulten and Sichel: measuring intangibles

“Measuring Capital and Technology: An Expanded Framework”, in Measuring Capital in the New Economy, 2005

- von Hippel: motivations of open-source software developers

"Open Source Software Projects as user Innovation Networks - No Manufacturer Required."
In Perspectives on Free and Open Source Software, 2007.

- Greenstein and Nagel : measuring Apache servers as substitutes

“Digital Dark Matter and the Economic Contribution of Apache,” NBER Working Paper 2013

- Sichel and von Hippel: measure household innovation based on time spent doing it.

"Household Innovation and R&D: Bigger than You Think." Review of Income and Wealth.

Data Development Questions

- How much is created each year? (flow measure)
- How much open-source software is in use? (stock measure)
- Who creates it? (Sectors: Business, Government, Academia, Households, Nonprofits, Foreign)
- What data can be used to develop a volume measure?
- What depreciation rates and deflators are appropriate?

Prototype for one Programming Language



OSS Definition	Registry	Repository (GitHub)
● Open Source Initiative (OSI) -Approved License ● Production Ready ● Release for Current Ecosystem	● Language-specific package managers (e.g., CRAN, PyPI) ● Contains metadata ● Continuous integration	● Commit Data (who, what, when) ● License ● Profile of contributors

- The registry data was collected using web harvest techniques.
- All CRAN and PyPI data as of July 2017, **14K R and Python packages** for analysis.

Language	R	Python
Package manager	CRAN	PyPI
Number of packages	13,719	164,836
Production ready	13,350	17,482
OSI-approved	13,143	15,043
Packages on GitHub (analysis)	4,358	9,773

[2] Robbins, C., G. Korkmaz, J. Calderón, D. Chen, A. Schroeder C. Kelling, S. Shipp, S. Keller. The Scope and Impact of Open Source Software as Intangible Capital: A Framework for Measurement with an Application Based on the Use of R Packages. *NBER Conference on Research on Income and Wealth*, Bethesda MD, March 15-16, 2019.

[3] Robbins, C., G. Korkmaz, J. Calderon, C. Kelling, S. Shipp, and S. Keller. The scope and impact of open source software: A framework for analysis and preliminary cost estimates. In *the 35th IARIW General Conference: The Digital Economy-Conceptual and Measurement Issues*, 2018.

Defining the Scope of OSS in the US

- Software that is published under an Open Source Initiative **OSI-approved license**.
 - Licenses establish permissions (e.g., use, inspect, modify, distribute, attribution) and limitations (e.g., liability, warranty).
 - Most common licenses are: MIT, Apache, GPL.

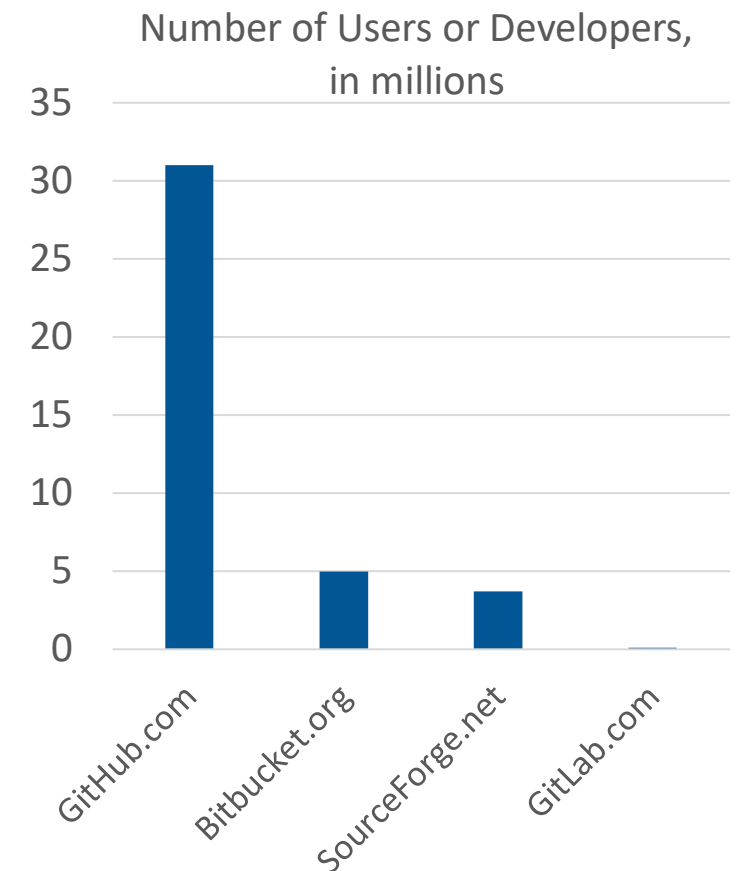
From prototype to scale-up:

1. **Packages for programming languages R and Python**

These are published codebases that are discoverable and installable through a registry and package manager.

2. **GitHub repositories**

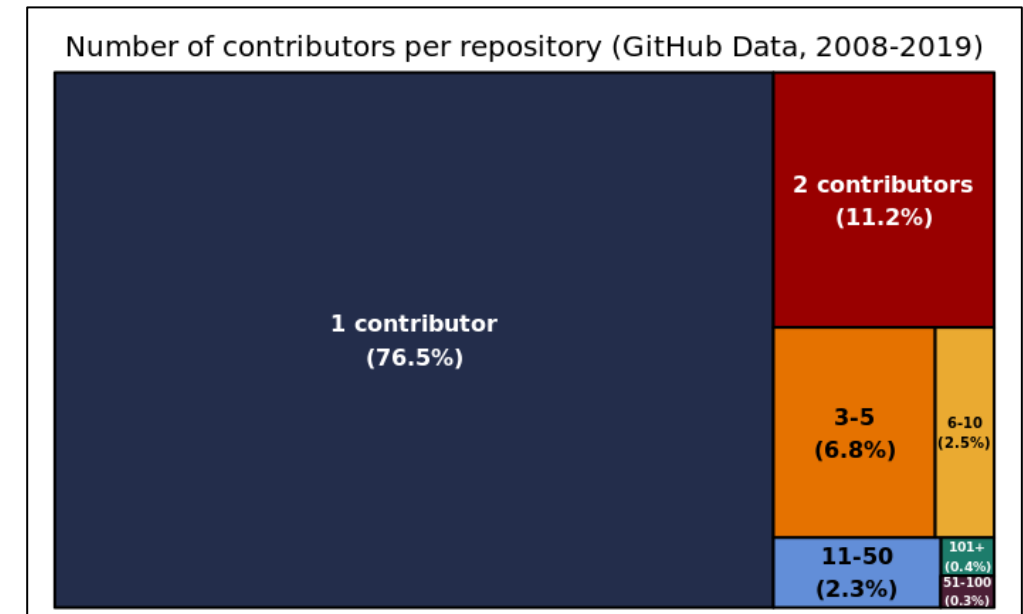
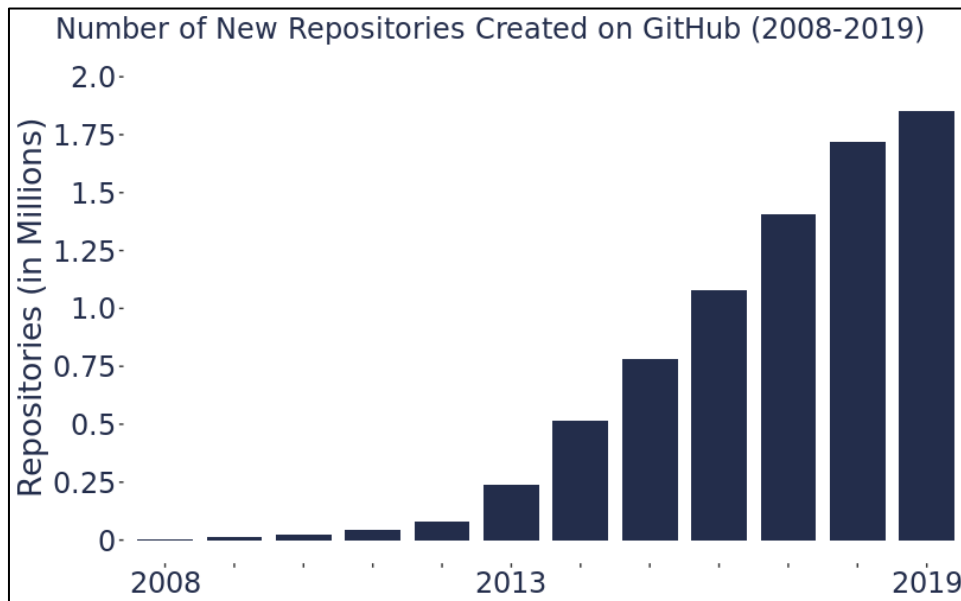
Repositories on GitHub, the world's largest remote hosting platform for Git version control.



Scale Up Data Collection: GitHub Repositories



- GHTorrent project data for additional user information (e.g., organization, company, location, email)
- Find public repositories with an OSI-approved license
- Collect information on development activity (e.g., commits, additions) and contributors using the GraphQL API.
- Obtained **7.75M repositories** (2009-2019) and **3.26M distinct contributors**



Source: UVA,
Korkmaz, Kramer,
Calderon, 2020.

Quantity/Volume of Output: How much is that?

Project length and complexity determine effort.

Software Cost Estimation: COCOMO II (Boehm, et al. 2000)

- Effort is a nonlinear function of complexity and lines of code
 - Code lines measured per project
 - Historical software project factors

$$\textit{Effort} = 2.4(KLOC)^{1.05}$$

$$\textit{Nominal development time} = 2.5(\textit{Effort})^{.38}$$

$$\textit{Labor cost} = \textit{Monthly wage} \times \textit{Nominal development time}$$

In Dollars, What Would that Imply?

Total resource cost = Resource cost (month) × Nominal development time

Labor costs: wage and salary plus nonwage compensation

Intermediate input costs

Taxes on production

Gross operating surplus

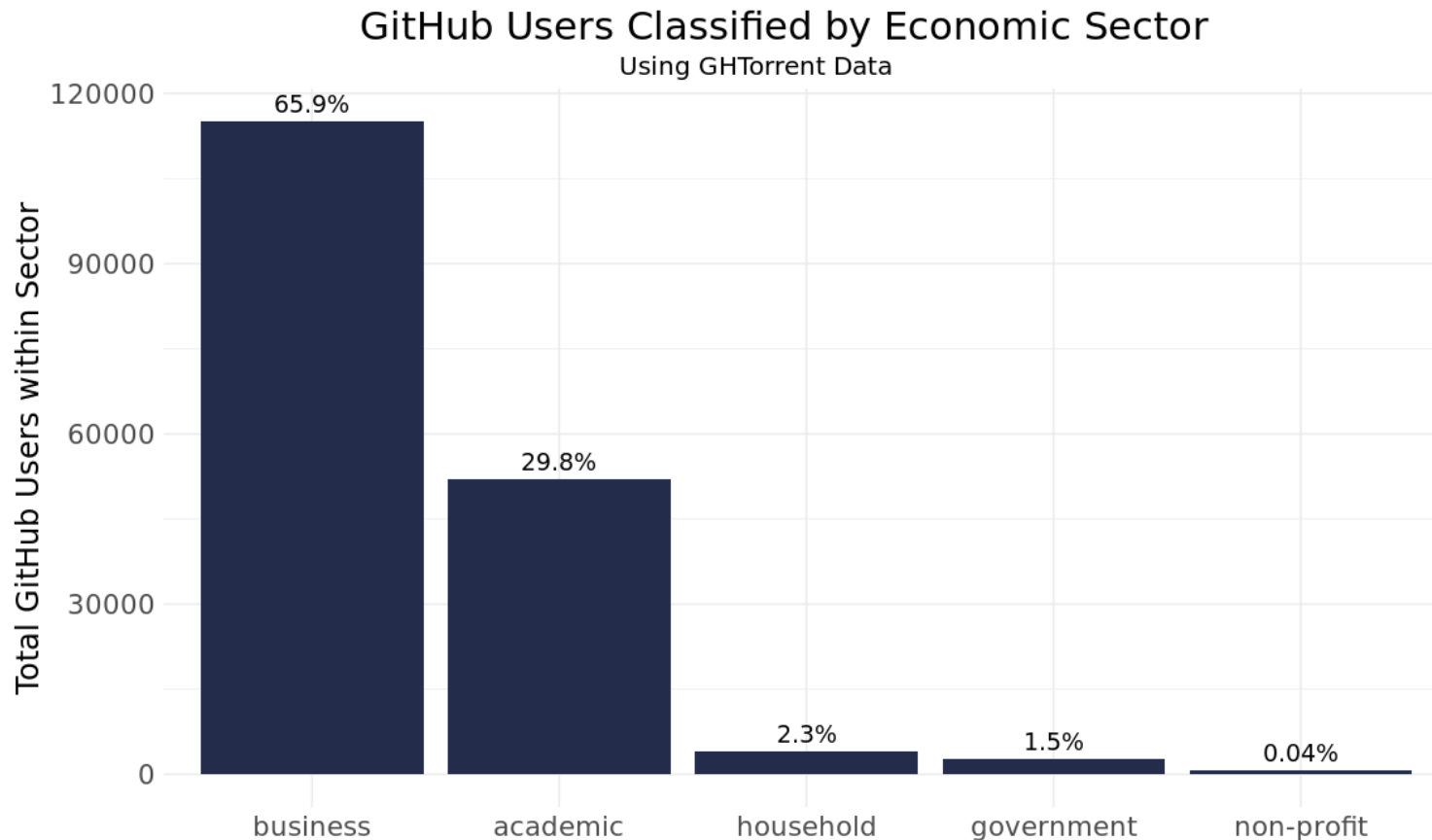
Prototype: 14K open-source packages registered in PyPI and CRAN and hosted on GitHub : **\$2.4 billion** (in 2017 dollars)

Scaleup: 7.75M GitHub repositories with OSI-approved licenses in 2019 investment total: 2.6M repos, cost based on lines added: **\$512 billion (2019)**

We can directly attribute \$33 billion to US contributors in 2019.

Sectoral Contributions

- Multiple data sources and methods used to estimate ***contribution of each sector*** taking into account collaborations across sectors



Use company field and emails in GHTorrent data to map developers to sectors
Mapped 20.4% of GHTorrent users to sectors.
12% of the total activity is captured

Source: UVA,
Korkmaz, Kramer,
Calderon, 2020.

Country-level Contributions

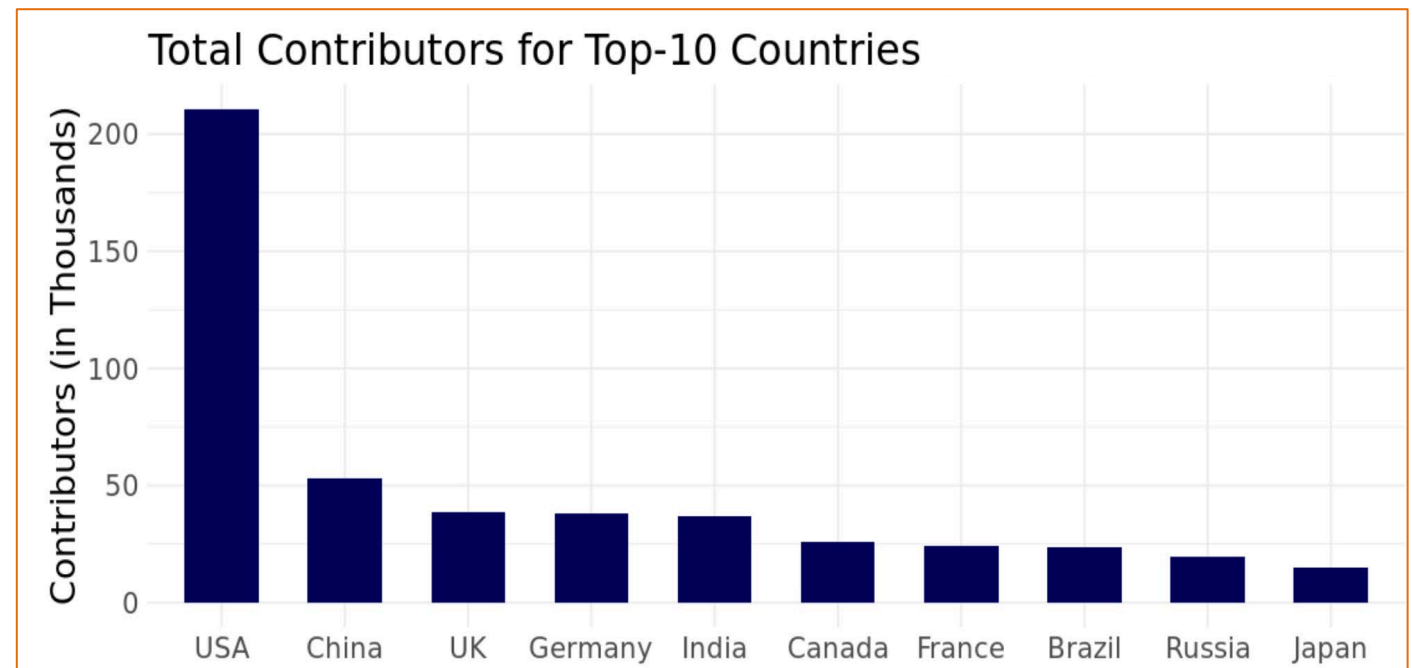
- **Contribution of each country** taking into account international collaborations (e.g., fractional counting).

Using self-reported **location information** in GHTorrent to map developers to countries (ISO-2C country codes, regular expressions, major cities, spelling fixes)

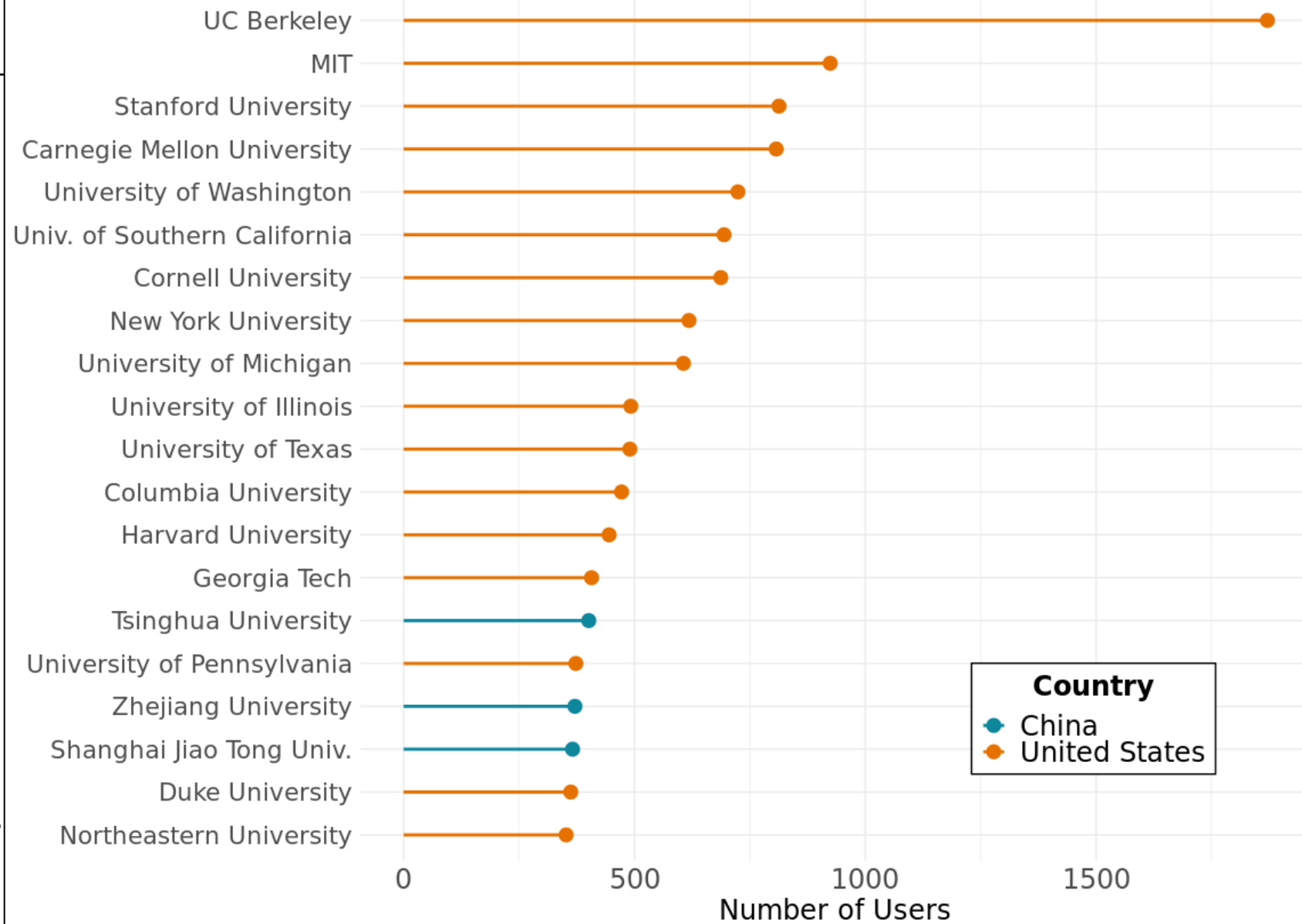
Mapped 19.7 % users in GHTorrent to countries. 33% of the total activity is captured.

US contributions are estimated as a third of the total contributions mapped to countries (35 %)

Source: UVA, Korkmaz, Kramer, Calderon, 2020.



Top-20 Universities by Number of Users



Source: UVA,
Korkmaz, Kramer,
Calderon, 2020.

Software Investment in Economic Output

	Private Sector			Public Sector			Household Sector	Rest of World
	Business	Other private nonprofits	Higher education	Higher education	Federal Government and FFRDCs	Non-federal government, ex. Higher Ed.		
Components of Software Investment								
Prepackaged								
Custom								
Proprietary								
Open Source (OSS)								
Own-account								
Proprietary								
Open Source (OSS)								

C. Robbins, G. Korkmaz, J. Calderón, D. Chen, A. Schroeder C. Kelling, S. Shipp, S. Keller 2019. *The Scope and Impact of Open Source Software as Intangible Capital: A Framework for Measurement with an Application Based on the use of R Packages*. National Bureau of Economic Research Conference on Research on Income and Wealth, Bethesda MD, https://www.nber.org/conf_papers/f111802/f111802.pdf

What we have learned so far

Quantities: Lines of Code and repositories

Contributors: by sector, academics will take more parsing

Countries: many contributors can be assigned

From Investment to Stock of Intangible Digital Assets

Next for us:

Annual Output/Volume based on own-account investment method, sum of costs

- Annual GitHub Volume: 2009-2019
- Price index: Own account software
- Depreciation rate: own account software

Measurement Questions

Can this approach translate to the creation of software in other economies (I/O ratios consistent)?

Does own-account software depreciate at the same or different rate than proprietary software?



National Center for Science and Engineering Statistics

<https://ncses.nsf.gov>