

Measuring the Scope and Impact of Open Source Software

Stephanie Shipp, Deputy Director and Professor
Social & Decision Analytics Division



UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE



NCSES

National Center for Science and Engineering Statistics

5 December 2019

Opportunities to observe and measure intangible inputs to innovation: Definitions, operationalization, and examples

Sallie Keller^{a,1}, Gizem Korkmaz^a, Carol Robbins^b, and Stephanie Shipp^a

^aSocial and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, Arlington, VA 22203; and ^bThe National Center for Science and Engineering Statistics, National Science Foundation, Alexandria, VA 22314

Edited by Katy Börner, Indiana University, Bloomington, IN, and accepted by Editorial Board Member Pablo G. Debenedetti August 25, 2018 (received for review February 16, 2018)

Measuring the value of intangibles is not easy, because they are critical but usually invisible components of the innovation process. Today, access to nonsurvey data sources, such as administrative data and repositories captured on web pages, opens opportunities to create intangibles based on new sources of information and capture intangible innovations in new ways. Intangibles include ownership of innovative property and human resources that make a company unique but are currently unmeasured. For example, intangibles represent the value of a company's databases and software, the tacit knowledge of their workers, and the investments in research and development (R&D) and design. Through two case studies, the challenges and pro-

including organizational innovation (3). Today, there are many examples of innovative outputs that are not sold in the market, such as open source software (OSS) and free online education (4). Furthermore, activities in the household sector, including inventions and social innovation (e.g., food delivery to poor rural children during the summer), are not included in summary data on innovation, because they are outside of the scope of business activity (5). There are many nonsurvey data sources created and used in the business and nonbusiness sectors that may provide signals that can lead to new measures of innovation.

Key Terminology

Open Source Software (OSS)

“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)

It is developed within and outside of the private sector

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

Current NCSES and other economic indicators do not measure the *value of open source software* outside the business sector.



Why Do We Care?

Open Source Software (OSS) are digital products, including those provided without direct payment

- OSS is **used across fields**; e.g., Google Chrome, Linux, R, Python, Wikipedia...
- OSS **supports research outputs**; e.g., peer reviewed publications, patents, startups, licenses ...

Innovation that is being created **outside of the business sector is not being measured**

- Missing a major contribution to economic growth



Challenge: Can the scope and impact of OSS be measured using publicly available data?

What is Currently Measured?

Components of Software Investment		Private Sector total in billions \$352.9*	Public Sector total in billions \$38.4	
Prepackaged		\$147.6	State and Local \$17.7	Federal \$10.7
Custom		\$141.1		
Own-account		64.3		

Source: BEA Intellectual Property Products Fixed Asset Tables (private) and Investment in Government Fixed Assets (Table 7.5B).

*Difference between Total and sum is a rounding error.

“The Open Source world is worth billions”

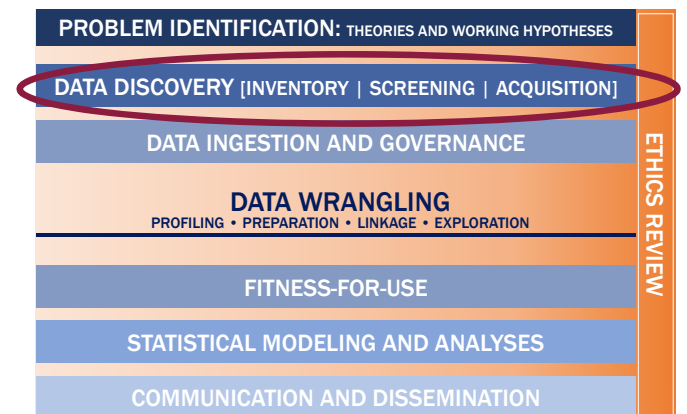
So what is the problem?

OSS created within universities, federal labs, and by individuals is not measured

Data Discovery

Desirable data dimensions for measuring OSS

- How much open source software is in use? (**stock measure**)
- How much is created each year? (**flow measure**)
- What types can be identified? (**categories**)
- Who creates it? (**sectors and collaborations**)
- Who uses it? (**attribution**)



Discovered Sources

- | | |
|-------------------|---------------------|
| • SourceForge.net | • GitHub |
| • OpenHub.net | • StackOverFlow.com |
| • Depsy.org | • OSalt.com |

Data Inventory and Acquisition

BLACKDUCK | Open Hub

Collection of OSS projects

- Development information and activity, contributors, users, commits
- Projects (677K); Organizations (698) and portfolio projects (2850), contributors

sourceforge

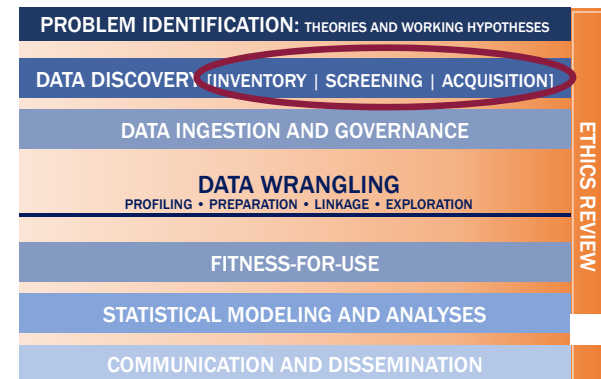
Collection of OSS activity

- Projects (~450,000) registered over 1999-2017
- Contributors (~350,000)
- User information (~291,782), project usage

depsy

Collection of OSS activity in Python and R

- 10,926 packages and 24,000 affiliated contributors
- Number of downloads, code reuse, citations in research papers, blog posts, tweets, views



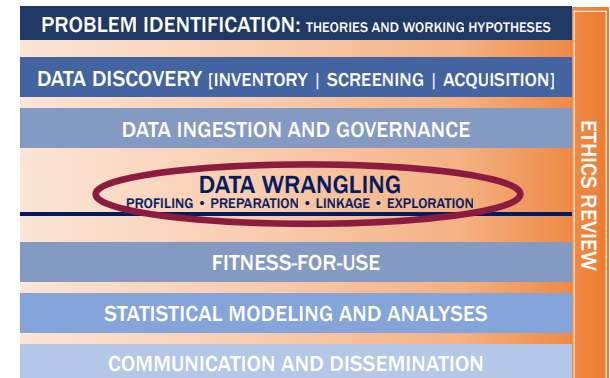
Data Profiling, Cleaning, Linking, & Exploring

Consider what pieces of the data may be able to be used in development models to measure impact and scope

Translate desirable data dimensions to variables that can be measured:

- How much open source software is in use? (**downloads**)
- How much is created each year? (**projects, lines of code, person hours**)
- What types can be identified? (**categories**)
- Who creates it? (**sectors, commits, networks**)
- Who uses it? (**citations, other developers**)

Can we find these variables in our data?



ggplot2

A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the ...



Tags

graphics phylogenetics

103 contributors

- RStudio
- Hadley Wickham
- Winston Chang
- kohske takahashi
- tidyverse

+ 98 more

View in API

Get badge

100 percentile impact overall

Compared to all research software on CRAN, based on relative downloads, software reuse, and citation.

Dependency PageRank

10.00 100 percentile

Measures how often this package is imported by CRAN and GitHub projects, based on its PageRank in the dependency network.

[Read more about what this number means.](#)



Downloads

6.3M 100 percentile

Based on latest downloads stats from CRAN.

Citations

1.7k 100 percentile

Based on term searches in ADS (0) and Europe PMC (1702)

[Read more about how we got this number.](#)

Reused by 9019 projects

Hmisc

Contains many functions useful for data analysis, high-level graphics, utility operations, functions...

ggmap

A collection of functions to visualize spatial data and models on top of static maps from various on...

rstan

User-facing R functions are provided by this package to parse, compile, test, estimate, and analyze ...

GGally

The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'gg...

lmerTest

Different kinds of tests for linear mixed effects models as implemented in 'lme4' package are provid...

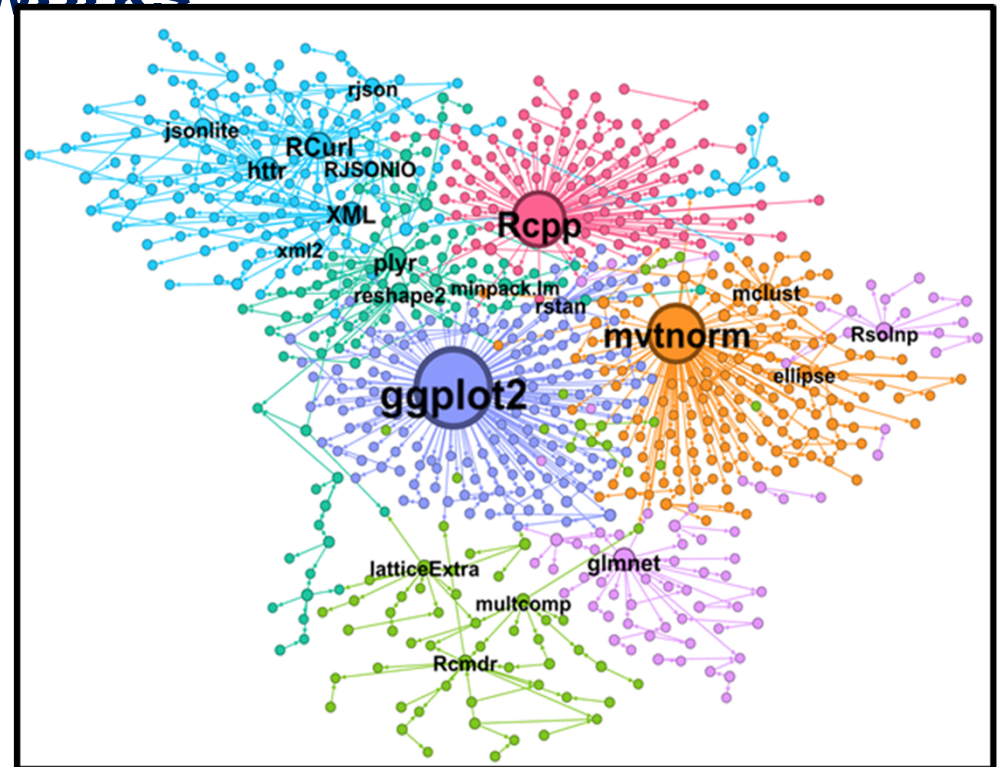
1859 ML_for_Hackers

Code accompanying the book "Machine Learning for Hackers"

Impact of OSS: Downloads, Reuse, and Dependency Networks

Package	2018 Downloads
Rcpp	3,519,510
rlang	2,893,889
stringi	2,610,184
stringr	2,511,011
ggplot2	2,495,315

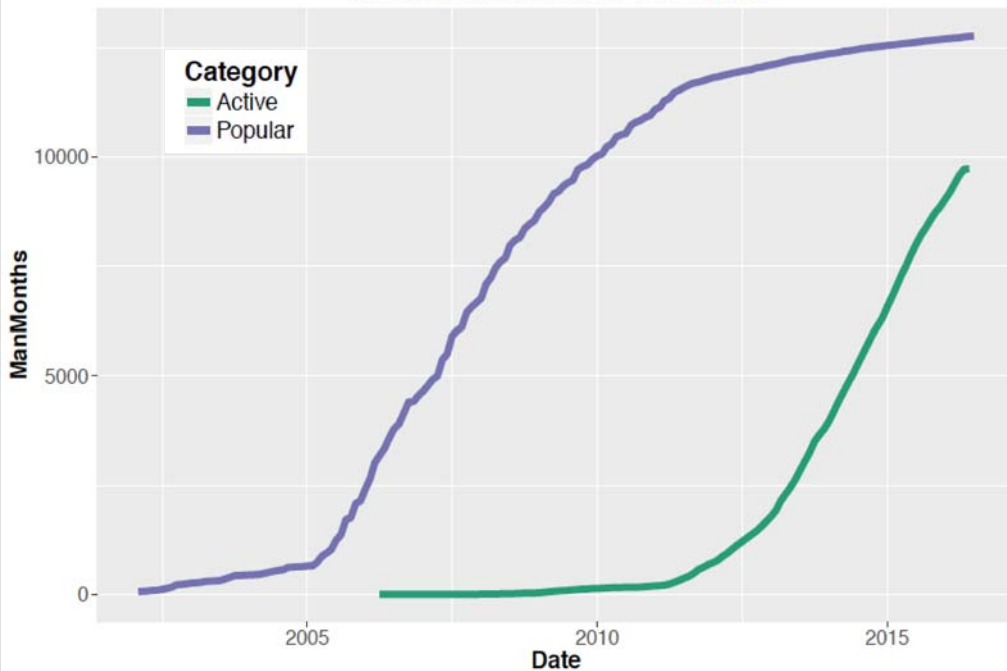
Package	Reuse
ggplot2	105,774
plyr	101,596
digest	99,774
stringr	98,086
colorspace	93,590



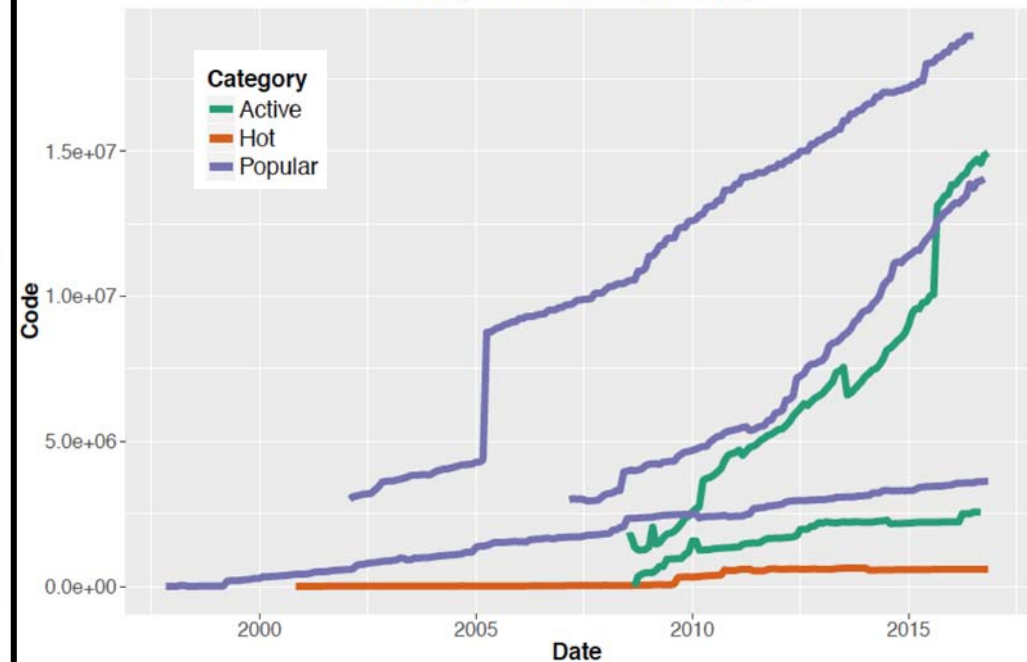
Dependency Networks

BLACKDUCK Open Hub

Total ManMonths vs. Date



Lines of Code vs. Date



BLACKDUCK Open Hub

Total Lines of Code



Total Commits



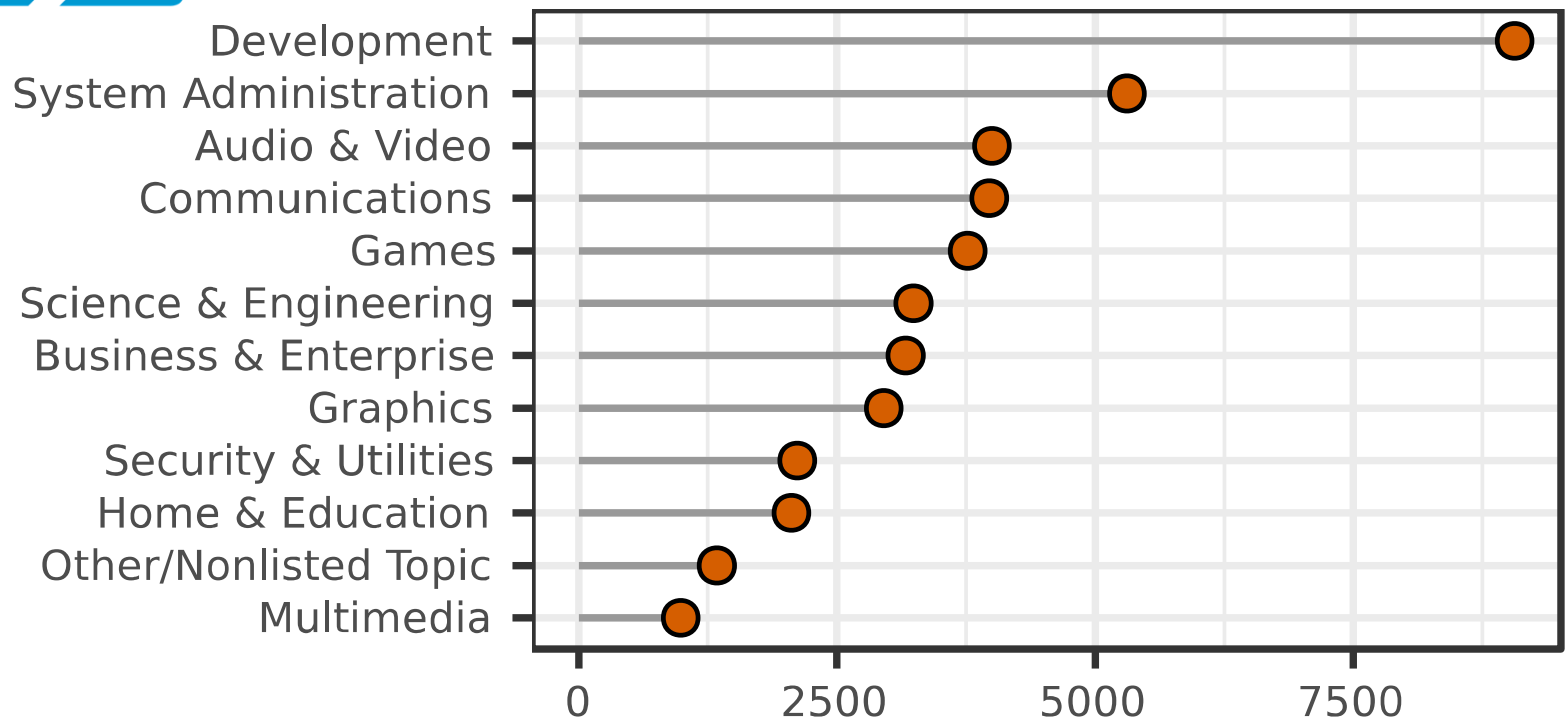
Total Contributors





Categories to Analyze Purpose

Median Downloads by OSS Category

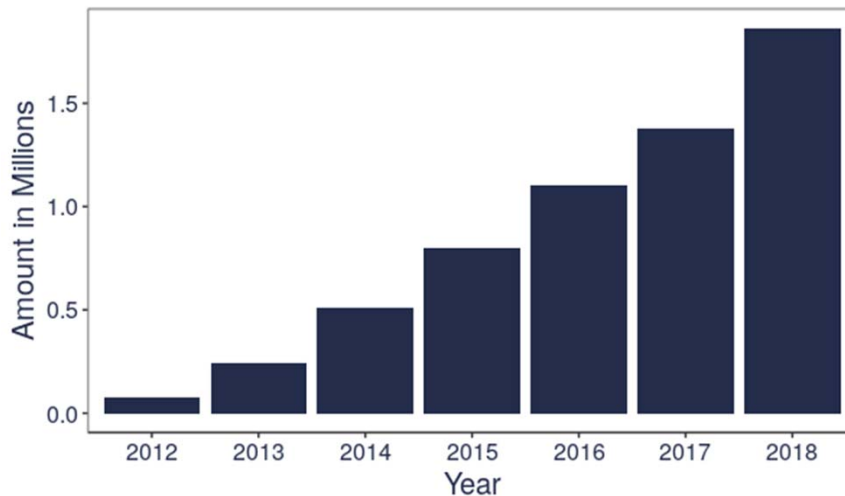


OSS Universe: Projects on GitHub

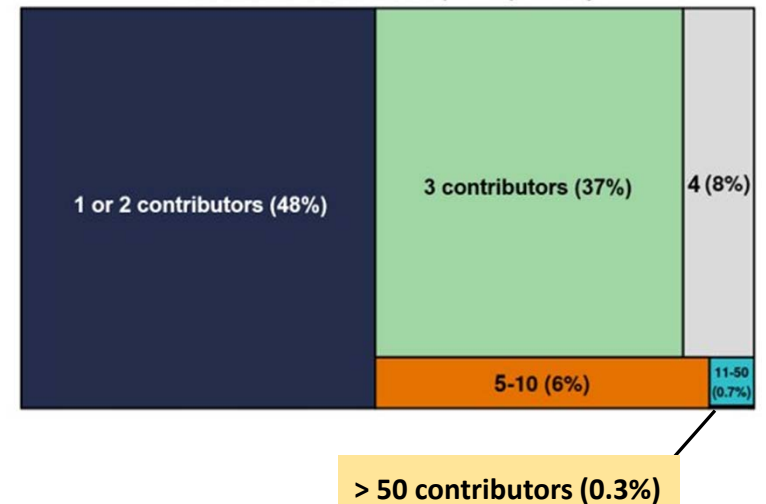


- 55.1M repositories with commits 2012-2018
- 7M repositories with OSI-approved licenses on GitHub as of July 2019
Of those, we analyzed **4.9M** repositories that have at least one commit
- There are **2.8M** unique OSS contributors

Amount of Repositories Created per Year in Millions

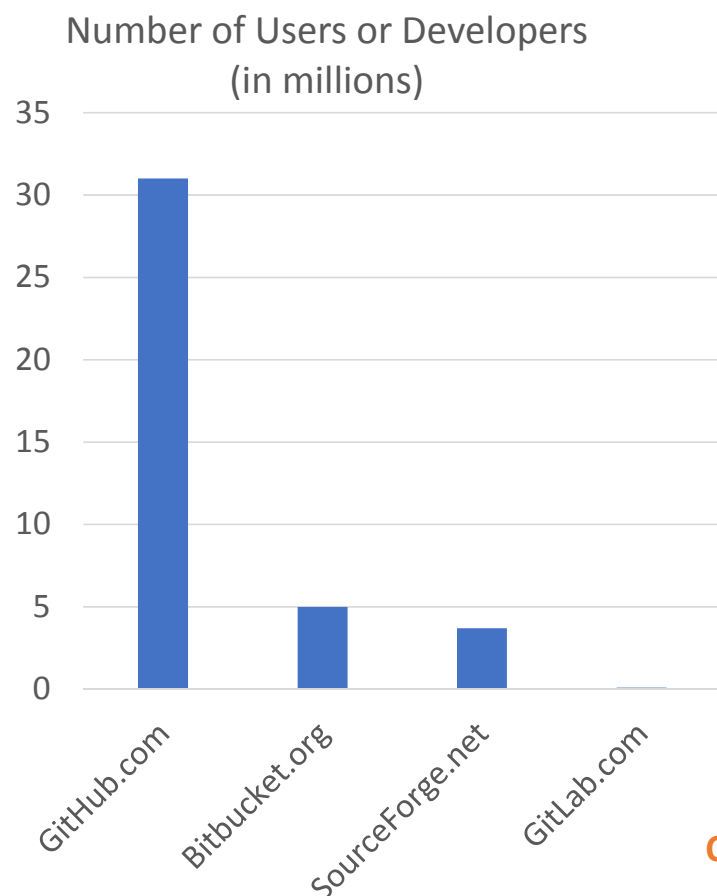


Number of contributors per repository



Most repositories have fewer than 5 contributors.

OSS Universe: Programming Languages



Language	R	Python
Package manager	CRAN	PyPI
Number of packages	13,719	164,836
OSI-approved & production ready	13,143	15,043
Packages on GitHub	4,407	11,016
Packages on GitHub (analysis)	4,358	9,773



Our data collection strategy combines data from multiple sources.

Value of OSS: Cost of Software Package Creation

- o Identify number of people involved each package's development
- o Estimate time spent on software development using **Kilo-lines of code (KLOC)**
- o Estimate **resource cost** with wage equivalent for 2017
 - Using average compensation for **computer programmers**
 - Occupation Employment Survey, Bureau of Labor Statistics
- o Estimate **non-wage costs**
 - Adapting BEA (Bureau Economic Analysis) and OECD (Organisation for Economic Co-operation and Development) methodologies



Cost of R and Python Packages developed on GitHub

Package Name	KLOC	Estimated Cost in Thousands of 2017\$
All packages	282,167.871	883,209
archivist	28488.639	4,169
CollessLike	15844.721	3,299
readtext	13888.309	3,130
ptwikiwords	11452.965	2,898
nasapower	10613.638	2,812

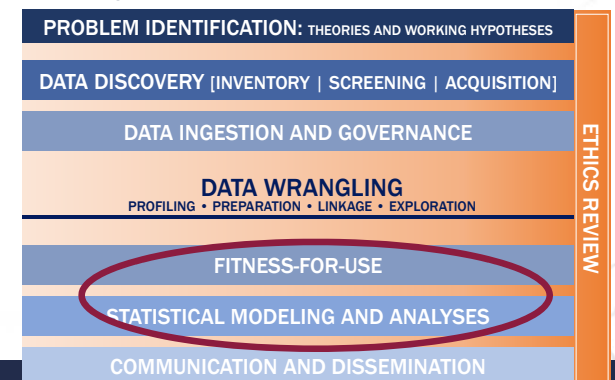
Package Name	KLOC	Estimated Cost in Thousands of 2017\$
All packages	611,601.568	1,560,374
libsass	50340.53	5,233
py3-ortools	37412.424	4,648
LSD-Bubble	15270.398	3,251
lotPy	14899.252	3,219
openquake.engine	13841.578	3,126

Returning to our research question: Can the scope and impact of OSS be measured using publicly available data?

Next steps - Build accurate and repeatable models to predict costs to produce OSS

- Cost estimation models are mathematical algorithms or parametric equations used to estimate the costs of a product or project
- Common attributes in software development cost models:
 - **Product attributes** (reliability, complexity, reusability)
 - **Platform attributes** (execution time, storage constraints, volatility)
 - **Personnel attributes** (capabilities of analysts and programmers; application, platform, language and toolset experiences)
 - **Project attributes** (use of software tools, multi-site development, required development schedule)

Fitness-for-Use - Evaluate data quality and utility for capturing these attributes



Communication & Dissemination

PEER-REVIEWED PUBLICATIONS

Keller, S.A.; G. Korkmaz; C. Robbins; and S. Shipp. 2018. "Opportunities to Observe and Measure Intangible Inputs to Innovation: Definitions, operationalization, and examples." *Proceedings of the National Academy of Sciences (PNAS)*. 115 (50), 12638-12645.

Korkmaz, G.; C. Kelling; C. Robbins; and S. Keller. 2018. "Modeling the Impact of R Packages Using Dependency and Contributor Networks." *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, pp. 511-514. IEEE.

CONFERENCE PRESENTATIONS/PAPERS

- National Bureau of Economic Research (NBER) Conference on Research in Income and Wealth (CRIW) Conference: Big Data for 21st Century Economic Statistics, Bethesda, MD, Mar. 2019.
- 2019 Women in Data Science Conference, Charlottesville, VA, Mar. 2019.
- International Association for Research in Income and Wealth (IARIW) 35th General Conference: Innovation and the Digital Economy, Copenhagen, Denmark, Aug. 2018.
- IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM), Barcelona, Spain, Aug. 2018.
- International Monetary Fund (IMF) Statistical Forum on Measuring Economic Welfare in the Digital Age: What and How? Washington D.C., Nov. 2018.
- NBER Conference on Research in Income and Wealth (CRIW) Pre-Conference: Big Data for 21st Century Economic Statistics, Cambridge, MA, Jul. 2018.
- Interagency Council on Statistical Policy (ICSP) Big Data Day, Committee on National Statistics (CNSTAT), Washington, DC, May 2018.
- Federal Committee on Statistical Methodology (FCSM) 2018 Research and Policy Conference, Washington DC, Mar. 2018.
- Arthur M. Sackler Colloquia of Sciences: Modeling and Visualizing Science and Technology Developments, Irvine, CA, Dec. 2017.

