Mapping the Research Software Ecosytem

Karthik Ram

UC Berkeley

James Howison

UT Austin

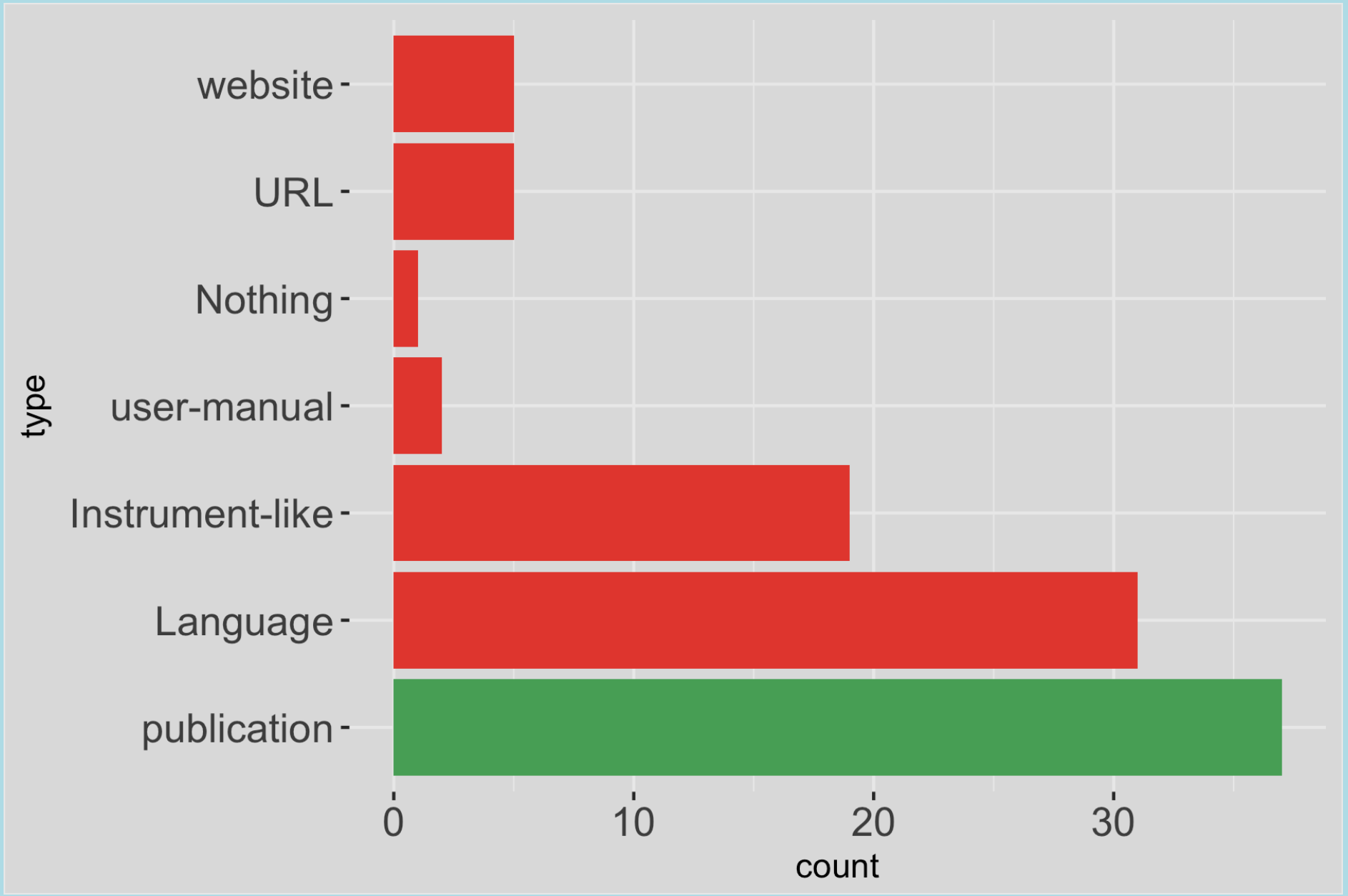
bit.ly/eoss-swmap

Use arrows or touch gestures to navigate slide deck

Research software

isn't a creditable

research activity



Howison & Bullard 2016

Formal citations: 31% - 43%

Informal mentions are the norm, even in high impact journals

Software is frequently inaccessible (15 - 29%)

Lack of visibility means that **incentives** to produce high-quality, widely shared, and collaboratively developed software **are lacking**

Prior work

Depsy


Heather Piwowar & Jason Priem

Software Heritage

Roberto di Cosmo

Libraries.io

Andrew Nesbitt

Scientific Software Network Map

Jim Herbsleb

Transitive Credit


Daniel S. Katz & Arfon Smith

World of Code

Chris Bogart

R-universe

dplyr

tidyverse/dplyr

A fast, consistent tool for working with data frame like objects, both in memory and out of memory.

Downloads
2.7 MM

100 Percentile Overall Impact
Compared to all research software, based on downloads, reuse and citation.

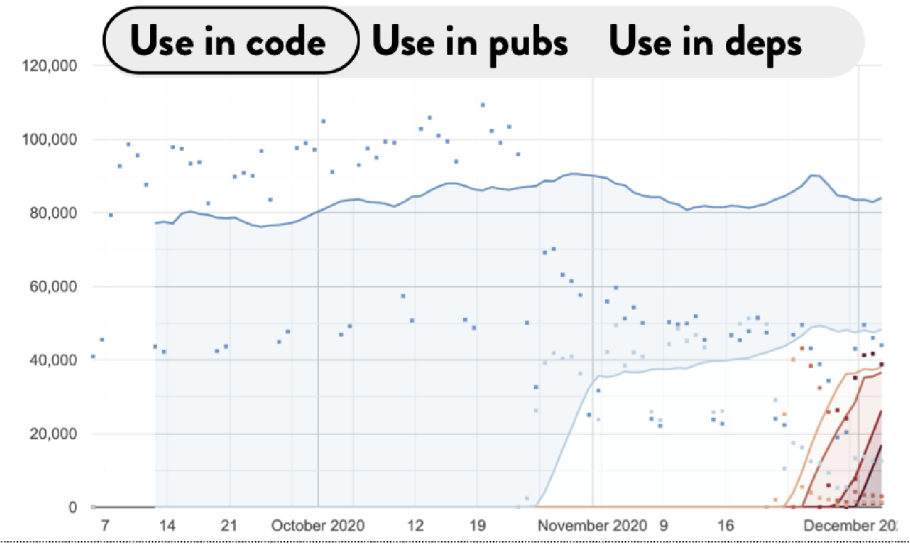
Dependency PageRank
10

Citation PageRank
90 Percentile Overall Impact

Summary

dplyr was originally released on Jan 15, 2014 and has since had **94** releases (~13.4 releases per year). The package has **107** contributors. Similar tools include data.table and pandas. The software is stable and widely used.

This is an infrastructure package and widely used as a dependency in many domains. Most frequently in **ecology**, **biomedical research**, **finance**.



Citations in context

Analysis was carried out in R version 3.5.2 (R Core Team, 2017) using the packages dplyr (Wickham et al., 2018) and devtools (R Core Team, 2017). Results Experiment 1: Ant mortality in response In: **The bat coronavirus RmYN02 is characterized by a 6-nucleotide deletion**...

Subsequently, background adjustments were performed by using the dplyr package. Finally, we utilized log2 transformation to normalize the data using the limma package. In: **Large-scale machine learning-based phenotyping**...

and 43,427 other papers...

Used by

The software is used in areas such as **Cell Behavior**, **Genomics, Molecular Networks**; **Neurons and Cognition**; **Populations and Evolution**; **Quantitative Methods**

mixOmics Multivariate methods are well suited to large omics data sets where the number of variables (e.g. ge...

and 1024 other repositories...

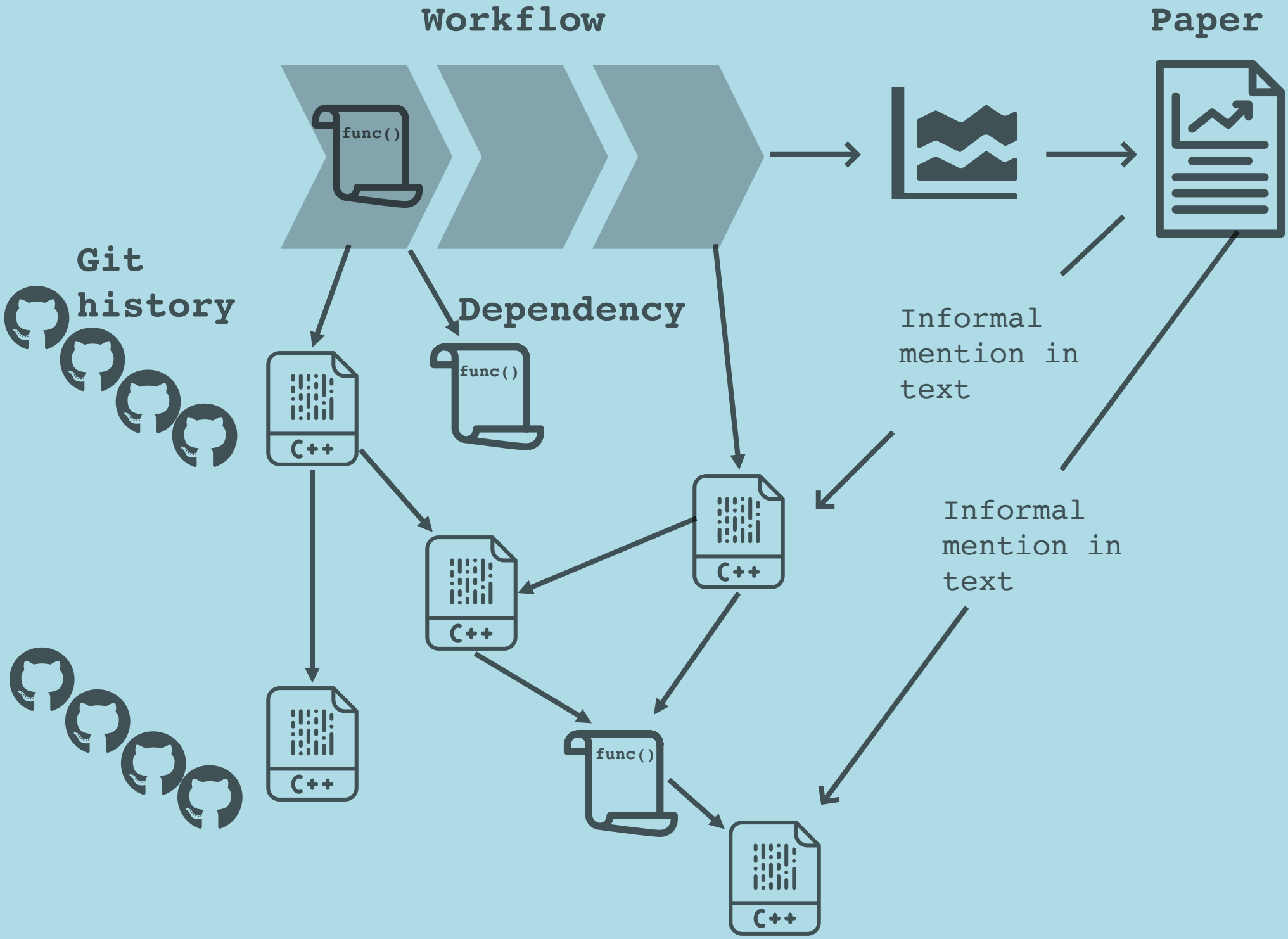
Used alongside

Based on an analysis of 1423 workflows, dplyr appears most commonly alongside tidyr, ggplot2, recipes, and broom.

Learn more about the ecosystem

How to cite

Wickham et al., (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686



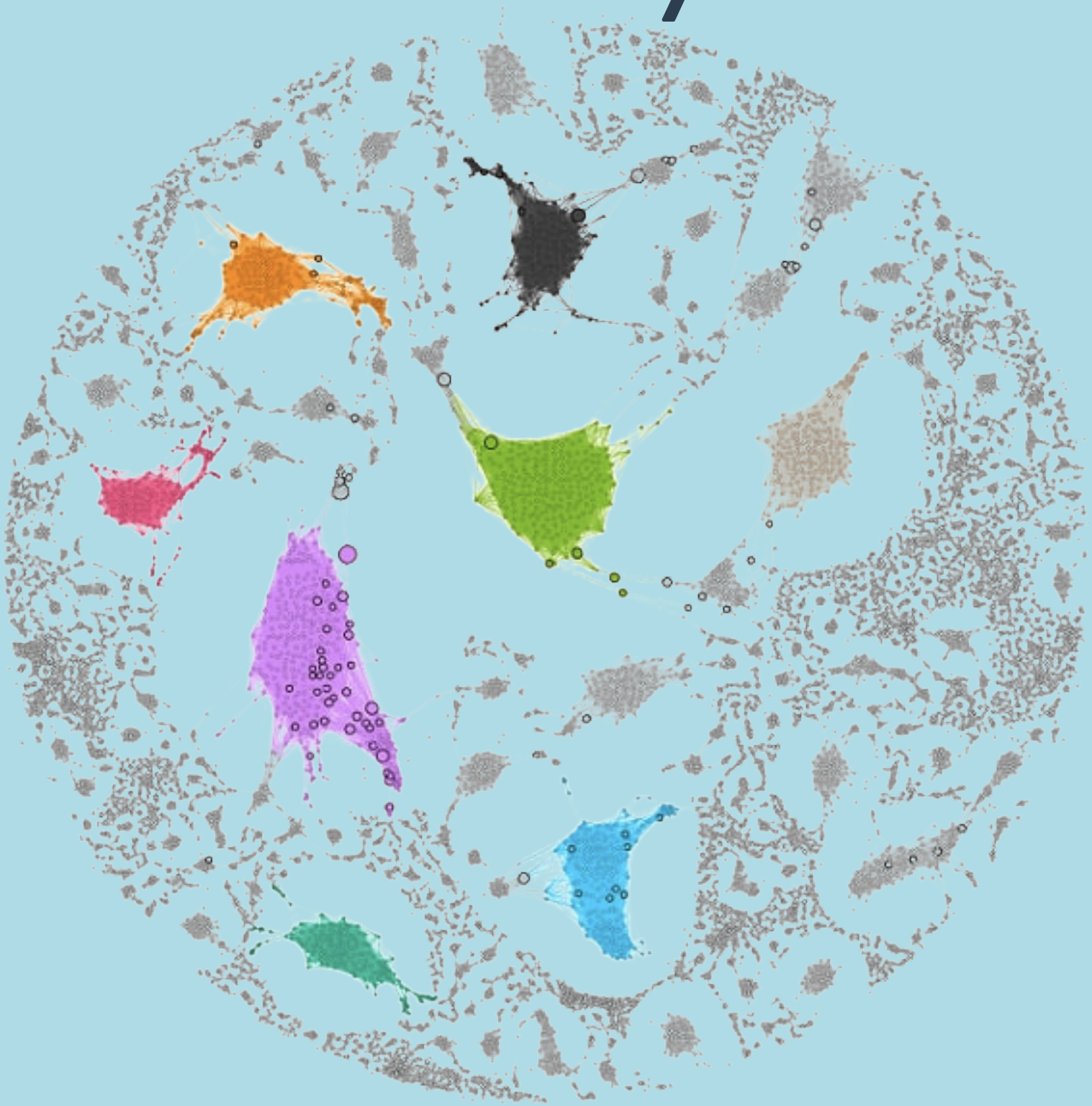
Paper text & metadata yield data on **topic**, **field**, **software use**

Package managers yield data on **dependencies**

Git repositories yield data on **history of development**

Time

Mapping the research software ecosystem



1

Which packages are increasingly used together in scientific workflows?

How might the map assist with how you know and interact with your project's upstream and downstream dependencies?

Can funding help make software more compatible?

bit.ly/eoss-swmap-q1

What groups of interdependent software are increasingly important for scientific fields?

How visible is their importance?

Can directed funding ensure the stability and maturity of critical dependencies and tool networks?

Does indirect usage do the work needed to demonstrate impact (with funders, with evaluators?)

bit.ly/eoss-swmap-q2

3

Which software components are seeing use outside their areas of original development?

Can funded interventions shore up interdisciplinary opportunities?

Are there “**leading**” and “**lagging**” fields? Can funded interventions bring lessons in achieving change within fields?

bit.ly/eoss-swmap-q3

4

How can we assess the weaknesses and opportunities in the ecosystem? Can project health data (like the CHAOSS project) be integrated to highlight strengths and weaknesses?

Which fields appear to be lagging? Can funded interventions bring lessons in achieving change?

bit.ly/eoss-swmap-q4

5

Can visibility of interdependencies motivate industry to provide pro-bono support to those building software crucial to science?

bit.ly/eoss-swmap-q5

6

What else do you want to do at an ecosystem level that this wouldn't help with?

E.g., Might we learn about how different structures of dependencies (e.g., proper hierarchies, hour-glass structures) affect the efficient flow of limited labor in science for bug reports, fixes, and improvements?

bit.ly/eoss-swmap-q6