The Science of Science Software

Some Updates

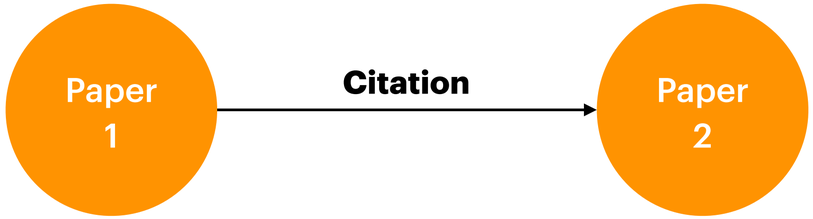
UNIVERSITY of WASHINGTON

SOS | 2024

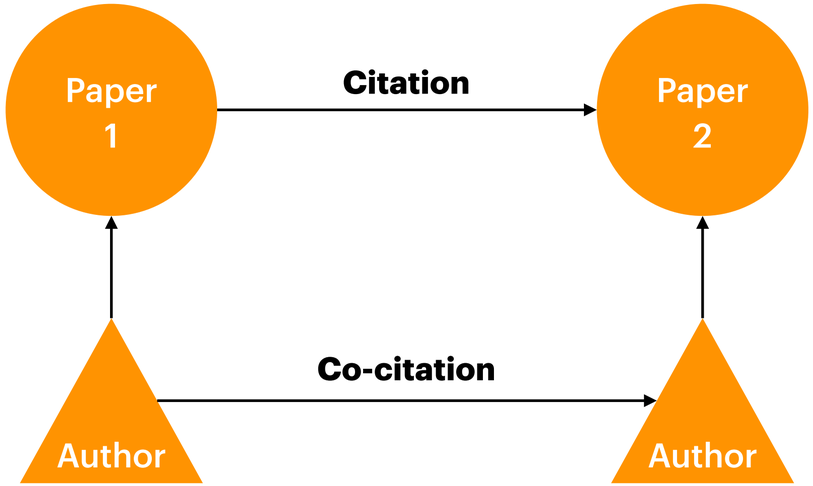
Overview

- **Empirical studies in SoSS**
  - Software Promises
  - Software Plans
  - Software Authors
- **Ongoing / Future Work**
  - Dependency Graphs
  - Citation Contexts

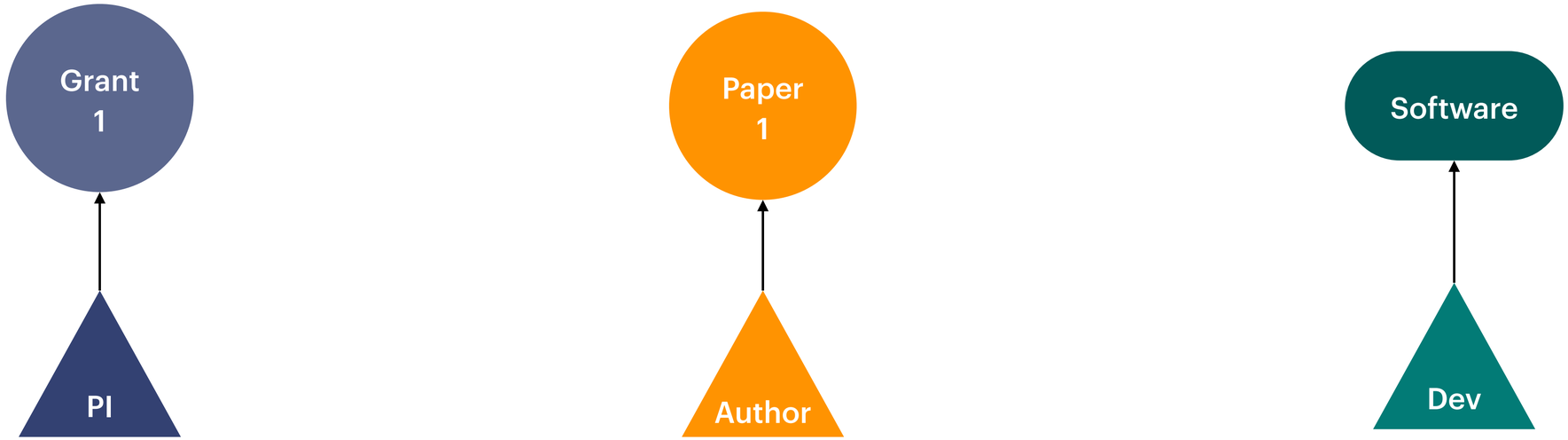
Science of Science



Science of Science

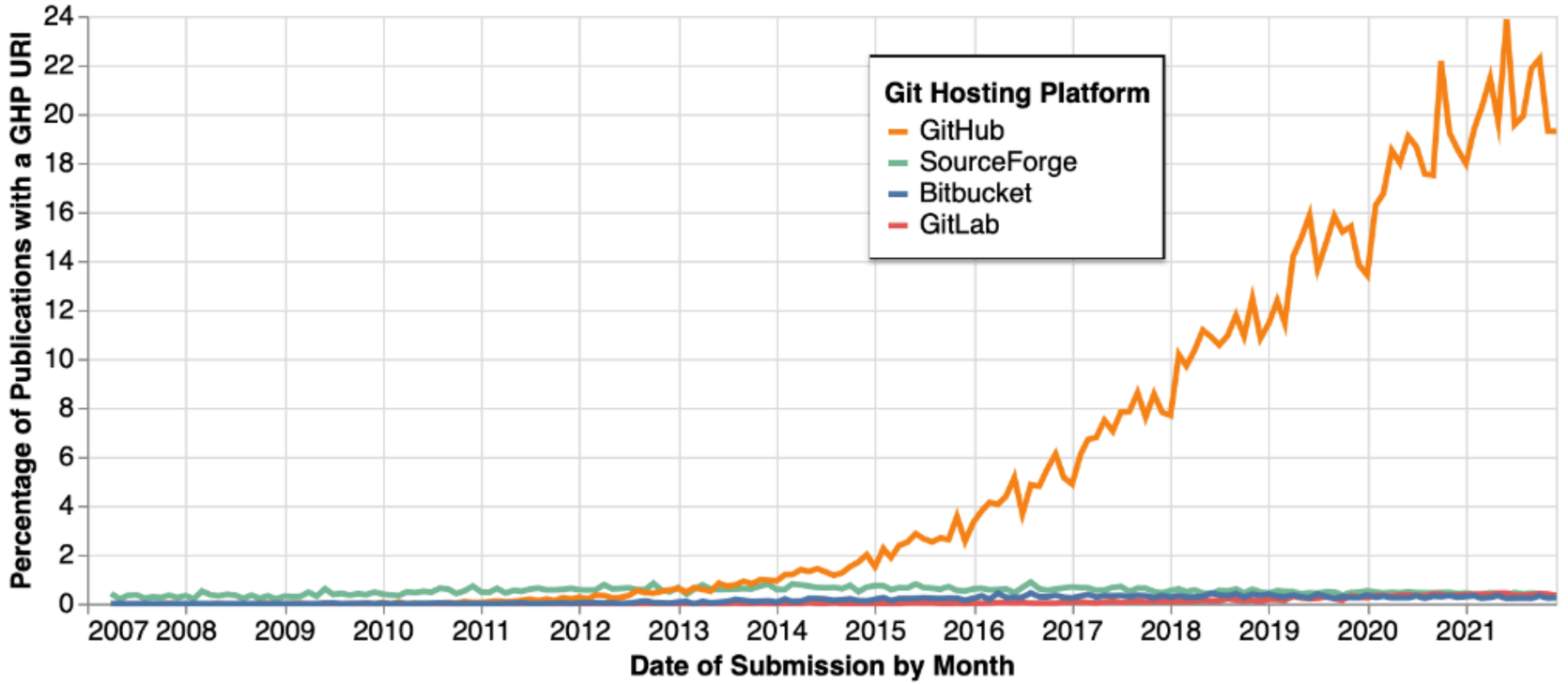


Science of Science



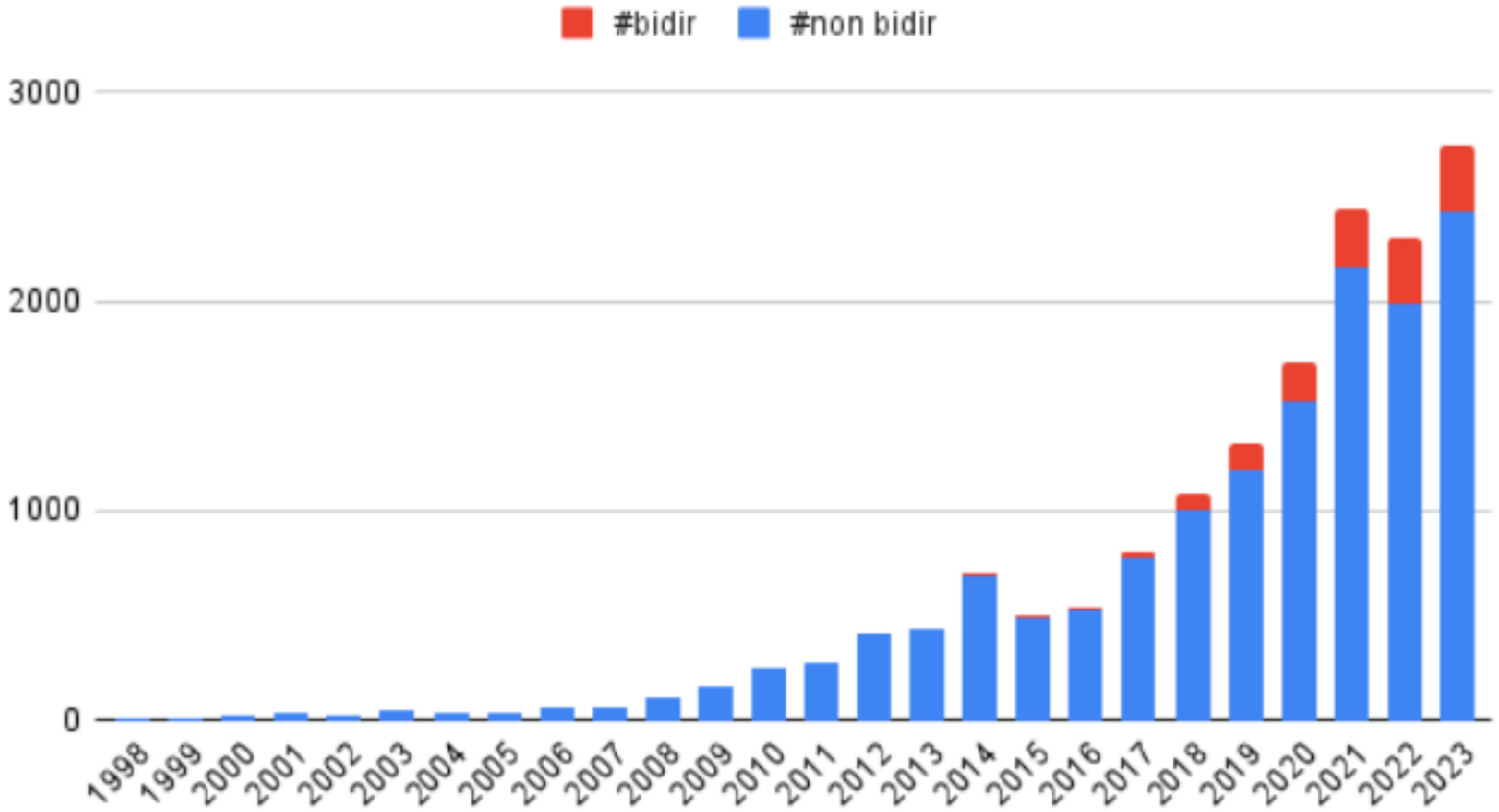
Science of Science Software

~25% of ArXiv papers link directly to public Git-backed repos



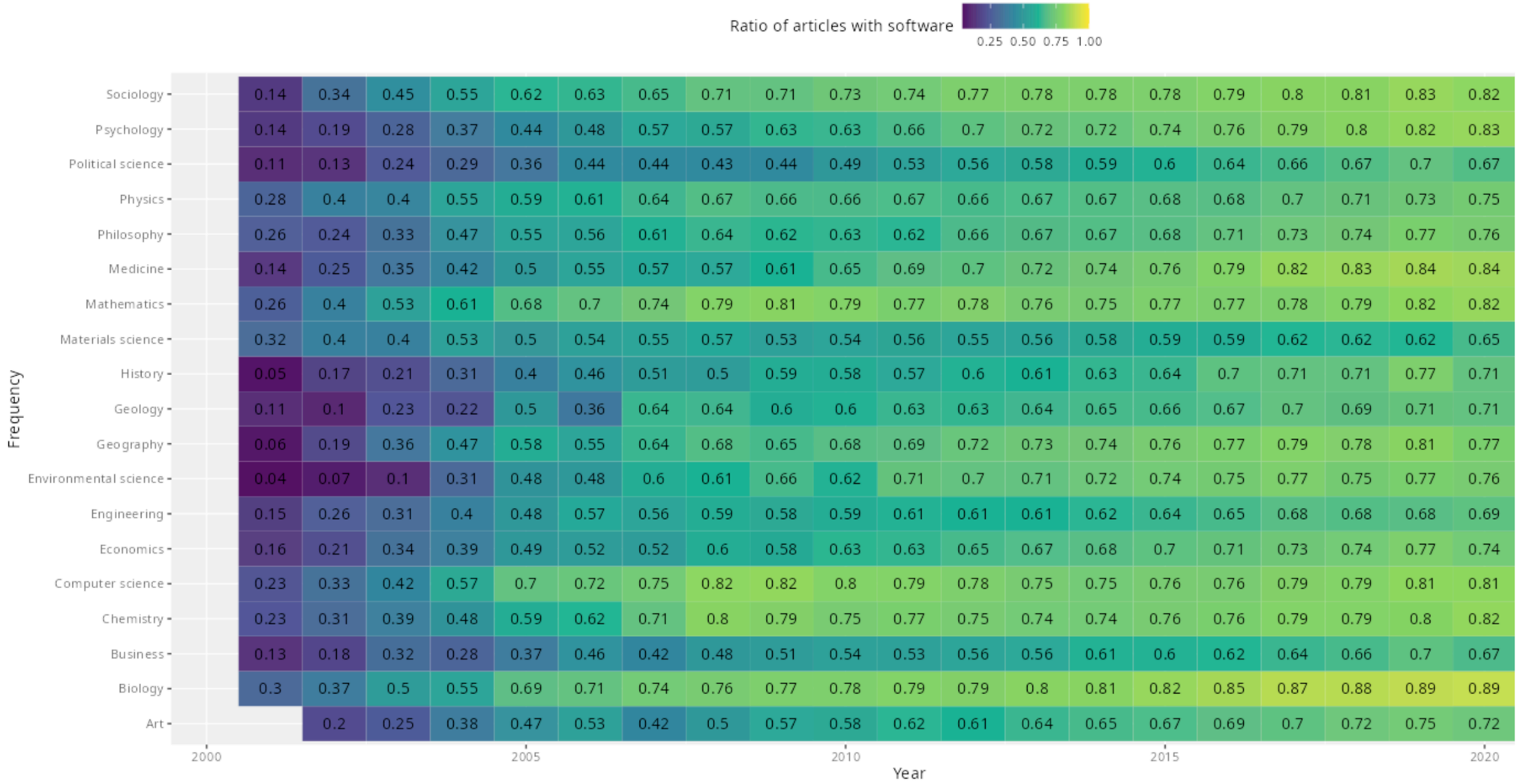
https://arxiv.org/abs/2208.04895

~1400 of ArXiv papers (in CS.se) have bidirectional link to Git-backed repos



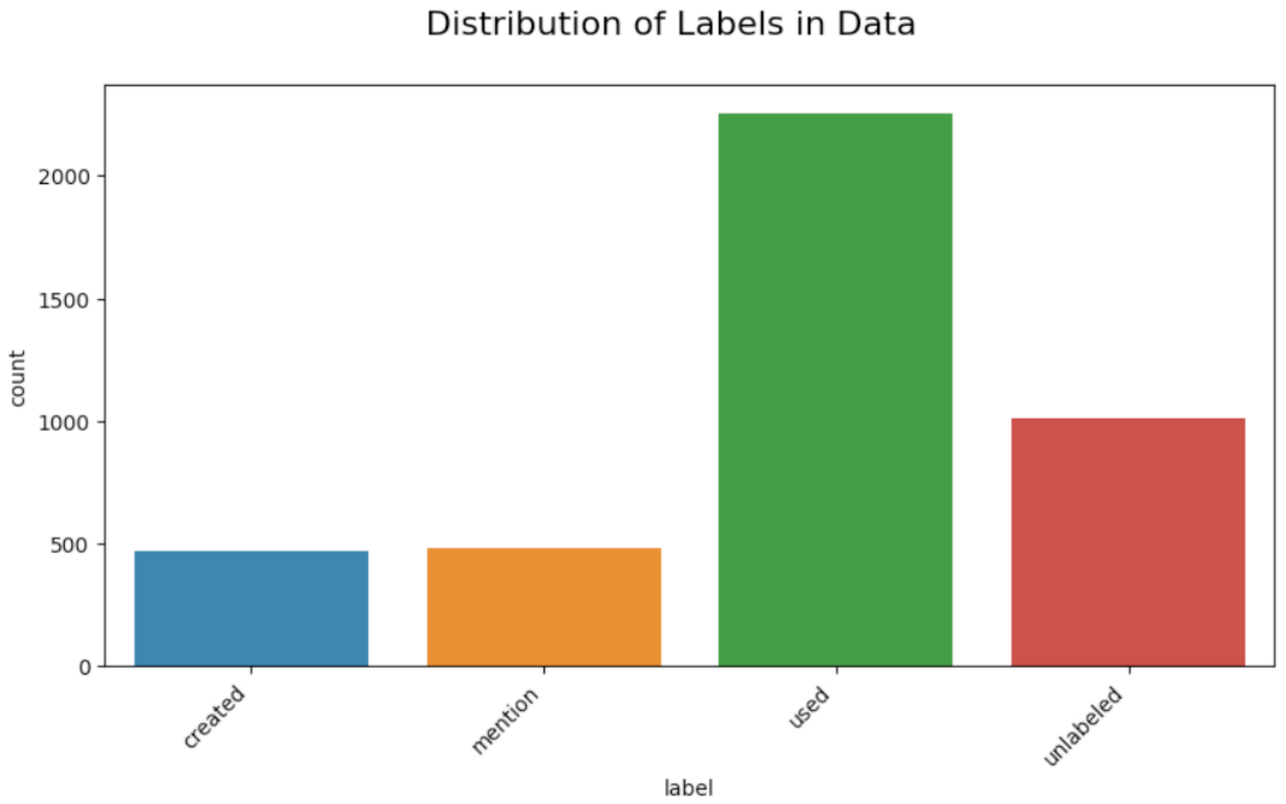
https://dgarijo.com/papers/msr_2024.pdf

By 2020 - Most disciplines cite or mention software



https://github.com/f-krueger/SoftwareImpactHackathon2023_DisciplinaryDifferences?tab=readme-ov-file

Distribution of software mentions or citations



https://github.com/karacolada/SoftwareImpactHackathon2023_SoftwareCitationIntent

Science Software Promises

Software Promises

Resources + Products

How many NSF awards produce software?

Award Data

NSF grant abstract and outcomes reports 2010-2012 = **~150k awards**

Approach

Use **embeddings** of a research grant's **proposal**...to predict software **produced**

Training Data

Repo -> Award = 1520 -> **446** explicit, unique, 'software' examples

Software Promises

Abstract

	model	accuracy	precision	recall	f1
0	tfidf-logit	0.674	0.674	0.674	0.673
1	transformer	0.636	0.608	0.697	0.649
2	semantic-logit	0.630	0.630	0.630	0.630
3	regex	0.516	0.515	0.516	0.514

Abstract + Outcomes

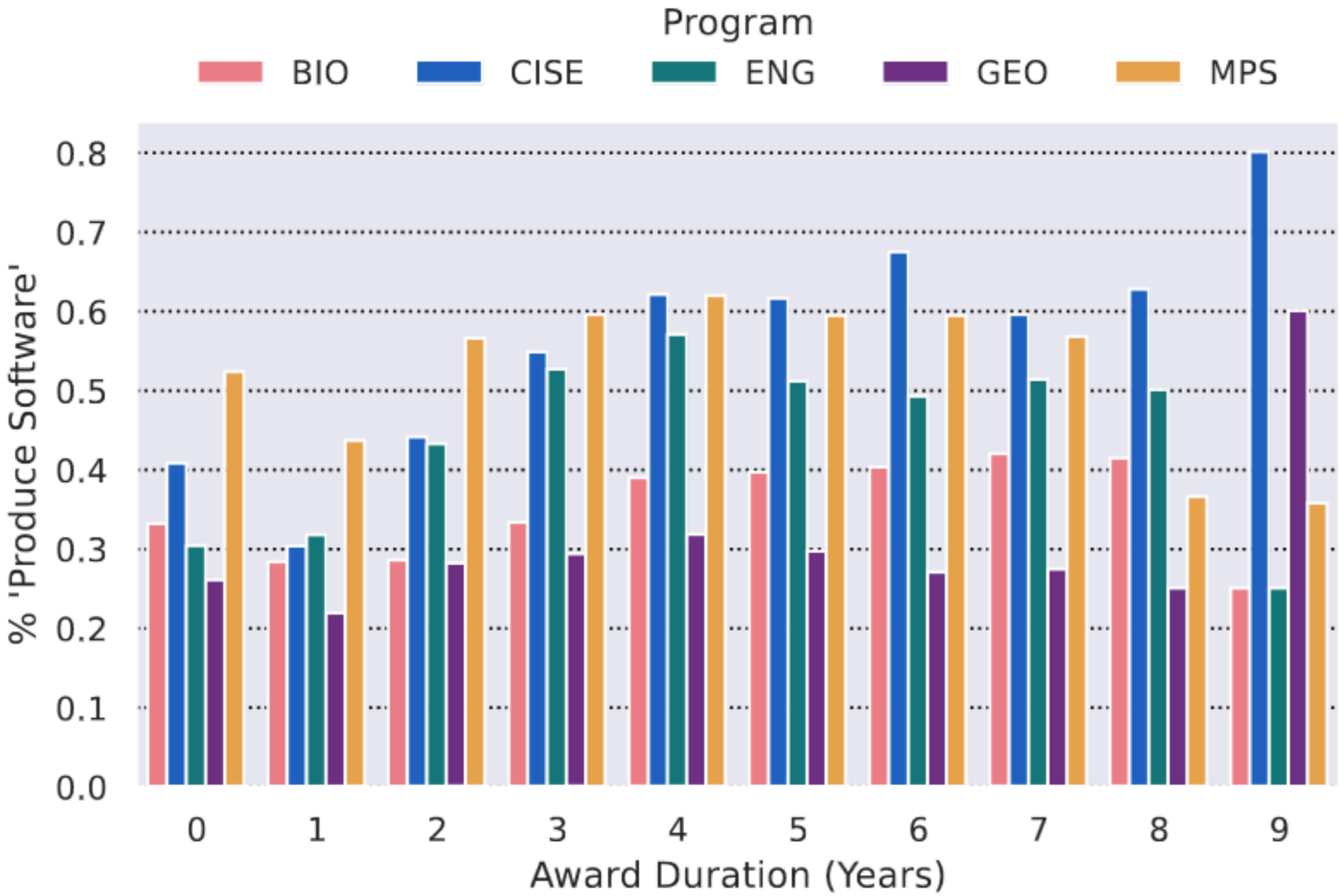
	model	accuracy	precision	recall	f1
0	tfidf-logit	0.745	0.745	0.745	0.745
1	transformer	0.673	0.638	0.771	0.698
2	semantic-logit	0.633	0.633	0.633	0.632
3	regex	0.510	0.507	0.510	0.482

Software Promises

	Program	# Awards	# Software	% Software
0	MPS	32885	19178	0.583184
1	CISE	24633	13274	0.538871
2	ENG	22900	11242	0.490917
3	GEO	17822	5142	0.288520
4	BIO	16990	6013	0.353914
5	EHR	13703	575	0.041962
6	SBE	13318	1966	0.147620
7	TIP	8597	4501	0.523555
8	OISE	2329	636	0.273079
9	OIA	498	123	0.246988

Software by NSF Directorate

Software Promises



Software Promises

Australian Research Council (Replication)

Training data

NSF corpus + 106 unique, linked, ARC repos

Grant data

ARC grant abstract 2010-2019 (no post-award data) = **~14K awards**

Software Promises

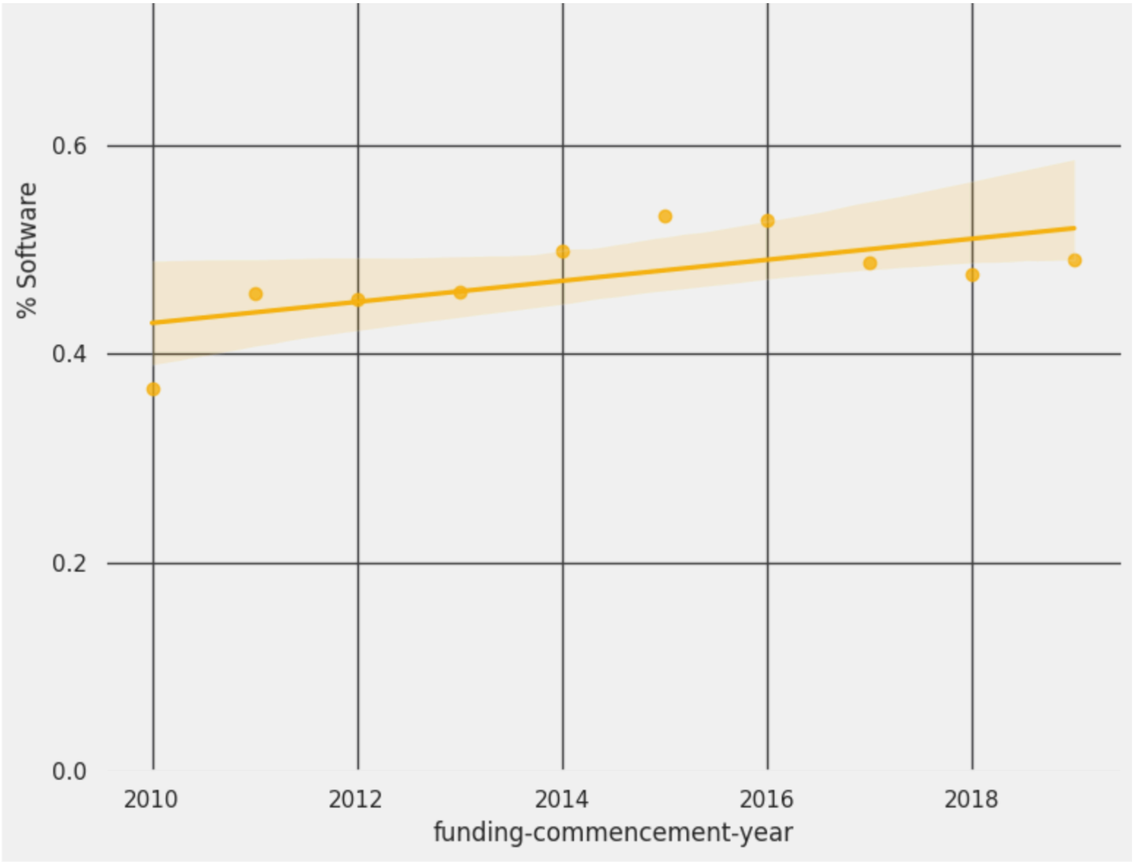
Abstract Only

Dataset	Model	Precision	Recall	F1
NSF	tfidf-logit	0.674	0.673	**0.6736**
ARC	tfidf-logit	0.815	0.696	**0.719**

47% of awards produce software

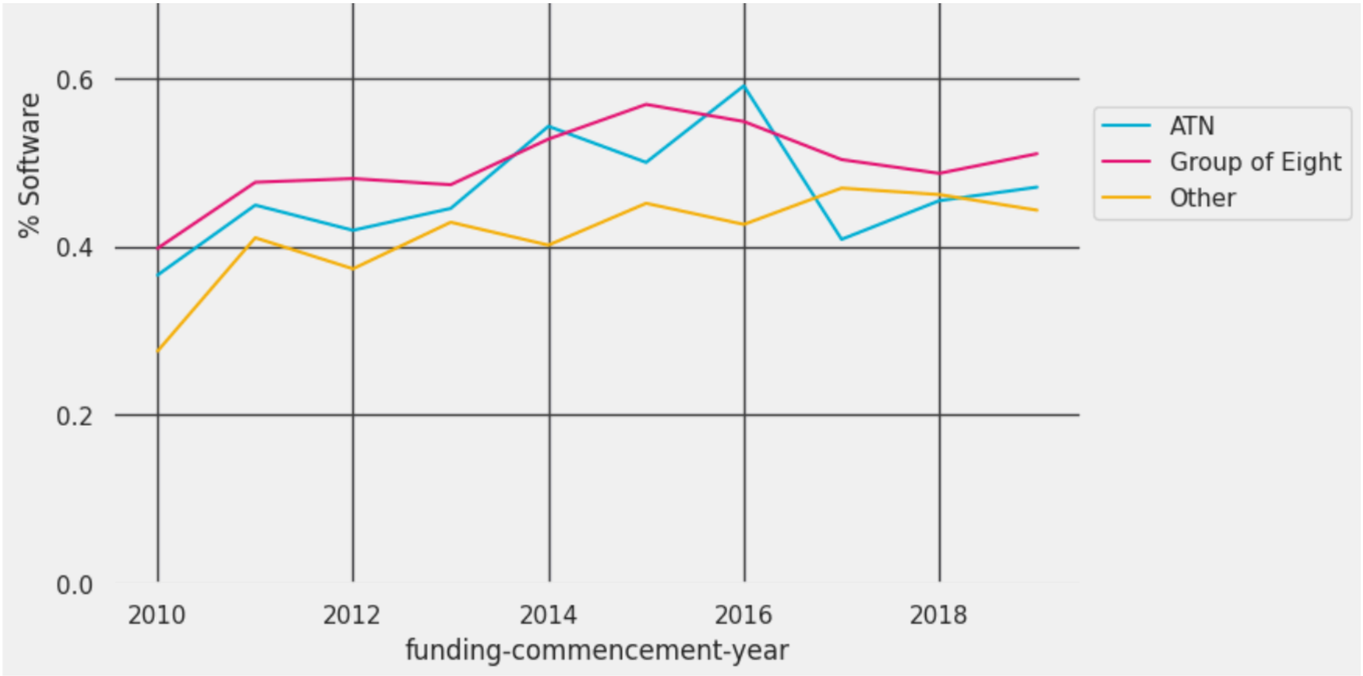
Software Promises

Awards Producing Software



Software Promises

Organisation Grouping	# Awards	# Software	% Software
Group of Eight	9203	4551	49.451266
Other	3285	1350	41.095890
ATN	1282	596	46.489860



Science Software Plans

Software Plans

Validation

How many NSF awardees from our sample (150K) produced software?

Question

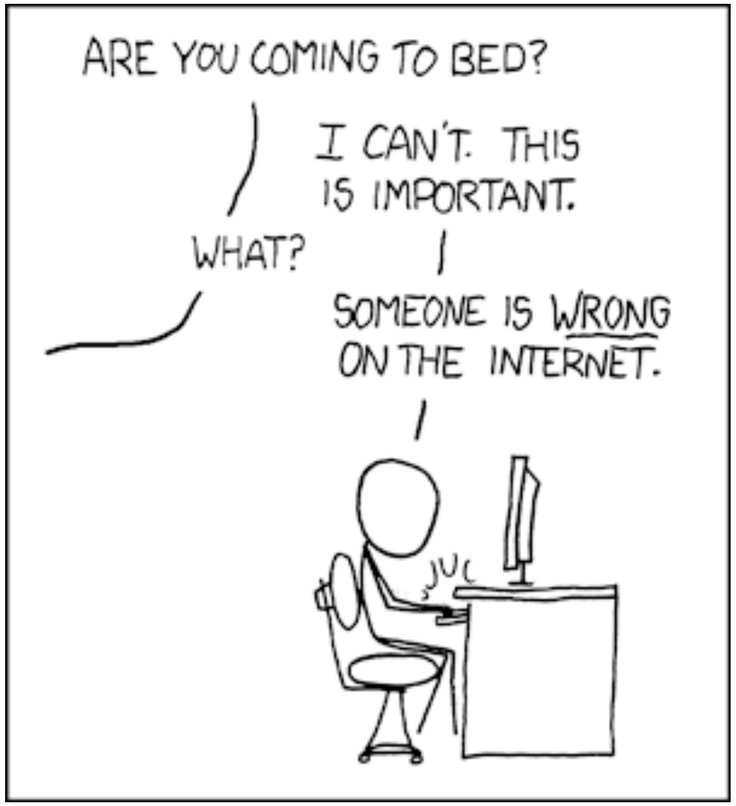
Did NSF awards plan to sustain their software (beyond grant) and if so, how?

Software Plans

Survey Experiments ...

"the best way to get the right answer on the internet is not to ask a question; it's to **post the wrong answer**."

Cunningham's Law



Software Plans

If we predict that an award **DID** produce software... our email to the PI explains that we predicted they **DID NOT ...**

We varied message (results in bold are statistically significant)...

- Subject line (NSF vs **Publicly-funded**)
- Identity (No identity vs **Scientists**)
- Prediction (Prediction vs **No-prediction**)

Software Plans

1629 responses (4.6% resp. rate)

892 produced software

.68 f1 - model performance 👎

Is software available?

... All Available: 41.37% (369)

... Partially available: 20.63% (184)

... Not available: 38.0% (339)

Software Plans

Why Not Available ...

- **Not-ready-for-public: 56.98% (298)**
- No utility: 36.9% (193)
- No time: 33.84% (177
- Too sensitive-data: 6.88% (36)
- Other: 12.81% (67)
- Own intellectual-property: 16.83% (88)

Software Plans

Plan to sustain software..?

... no plan: 33.87% (211)

... plan for some software: 33.71% (210)

... plan for all software: 32.42% (202)

Software Plans

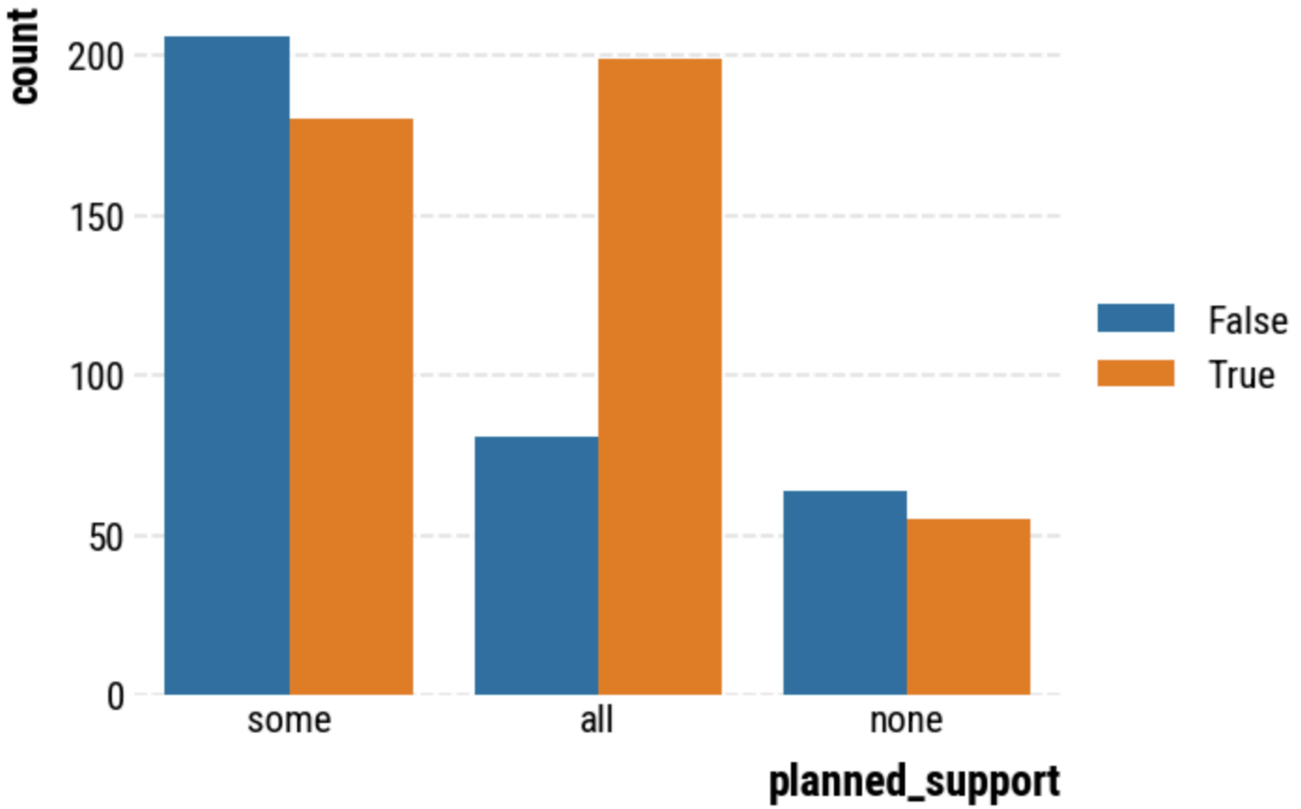
Did plan...

- Research: 51.77% (321)
- Used-by-others: 40.16% (249)
- Teaching: 16.13% (100)
- Other: 15.48% (96)
- Required by funding: 9.19% (57)

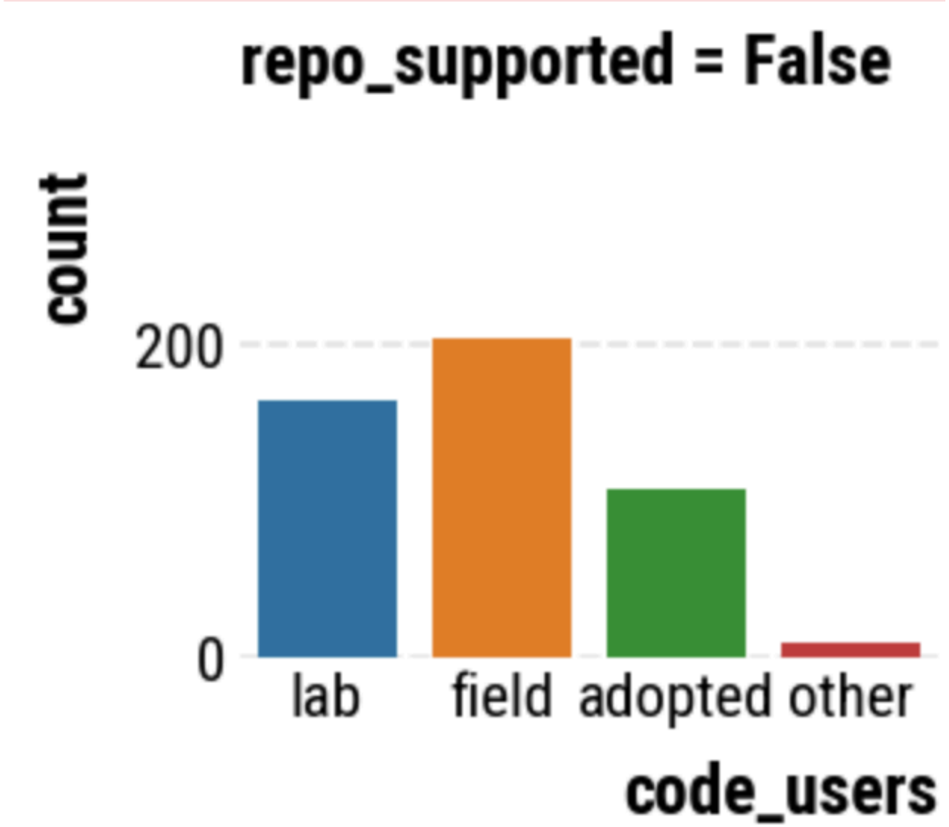
Did not plan...

- No-funding: 19.68% (122)
- no-time: 16.94% (105)
- no-research: 13.71% (85)
- no-use: 14.03% (87)
- no-teaching: 12.26% (76)
- no-credit: 14.19% (88)

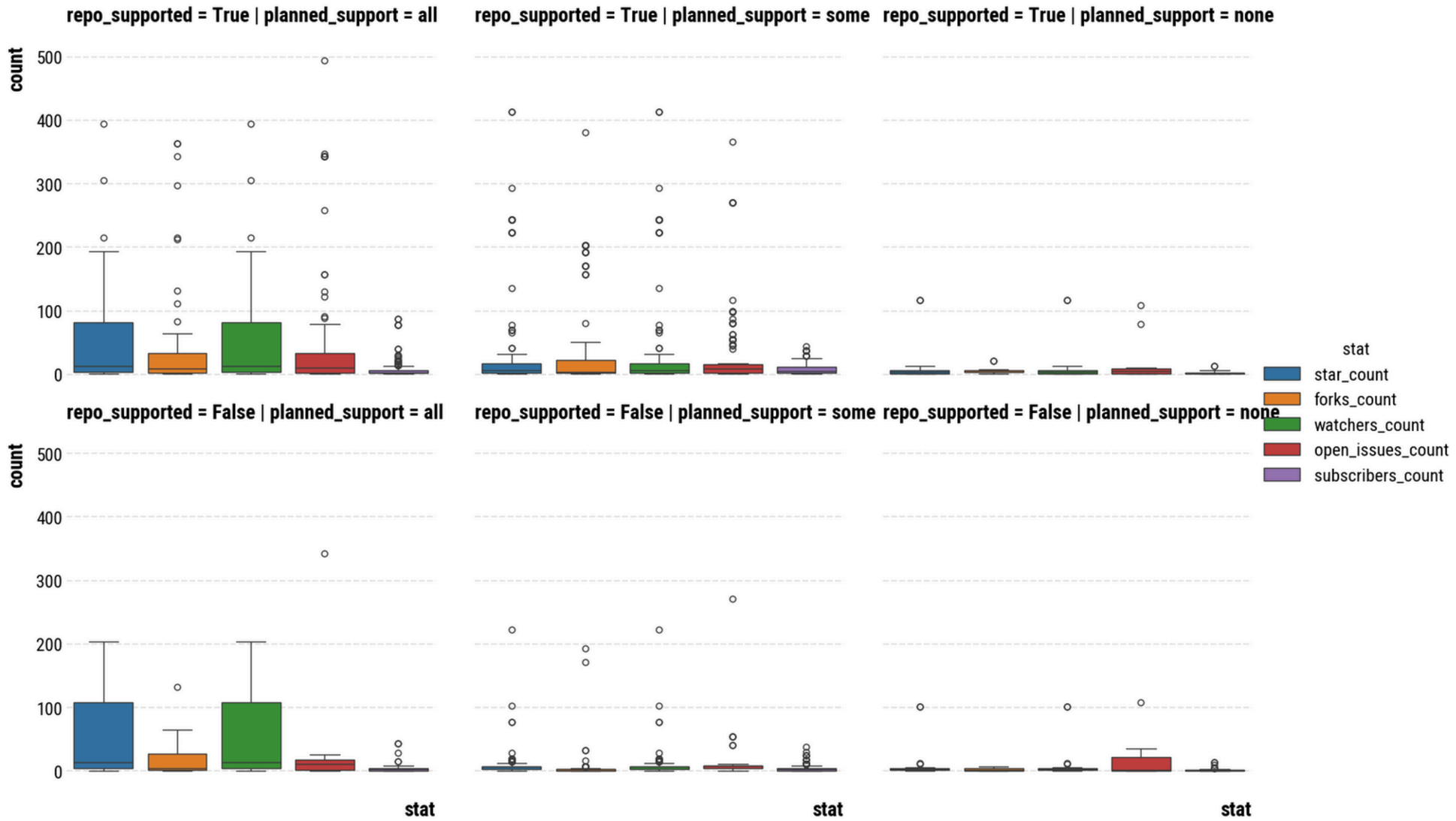
Software Plans



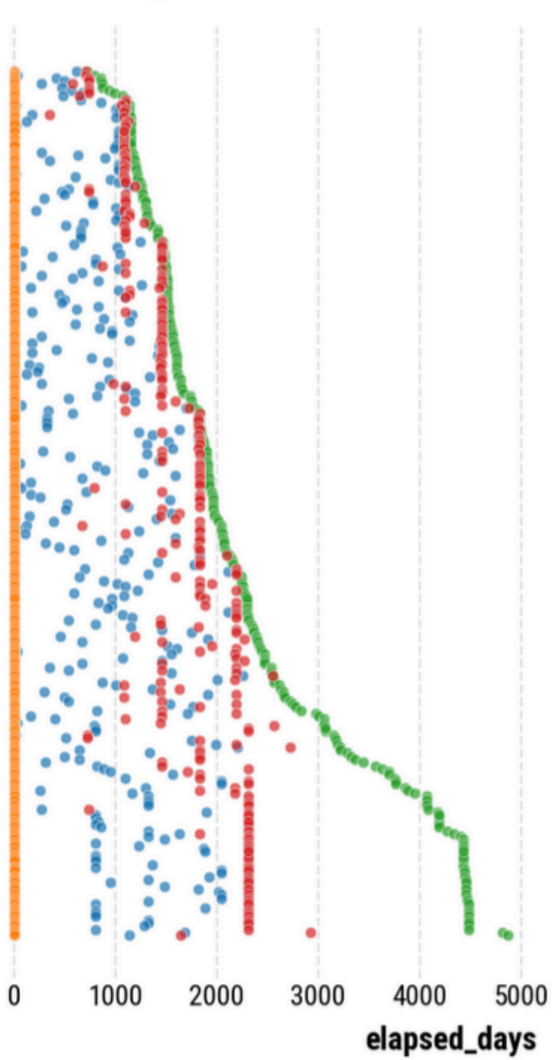
Software Plans



Software Plans



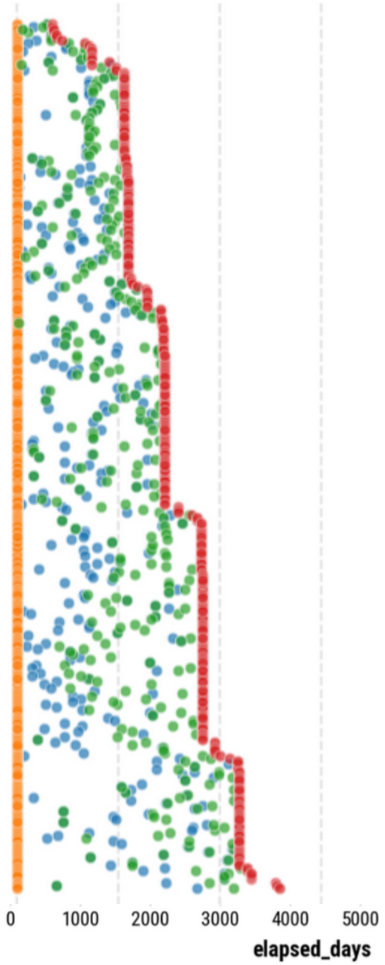
Software Plans



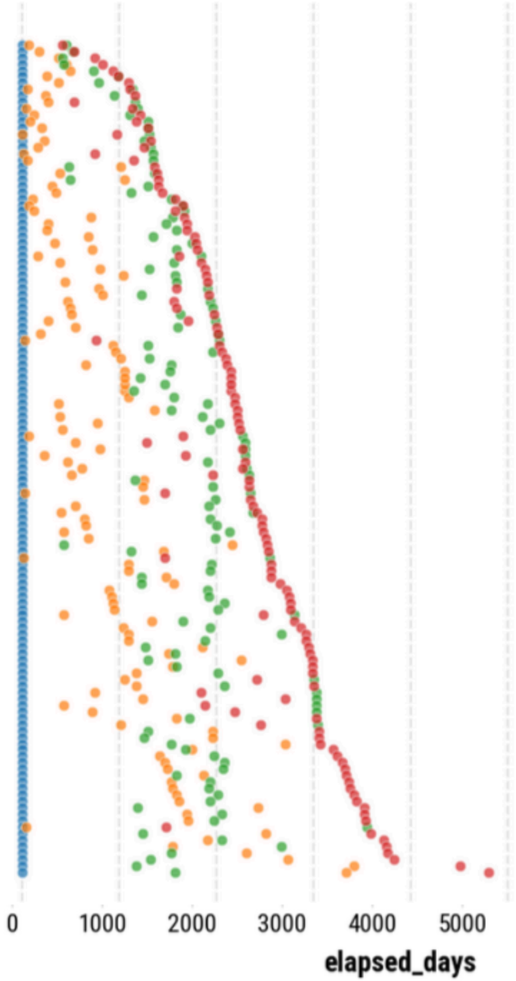
56% (n=552) made a commit after the grant ended

Software Plans

During



Before



Science Software Future Work

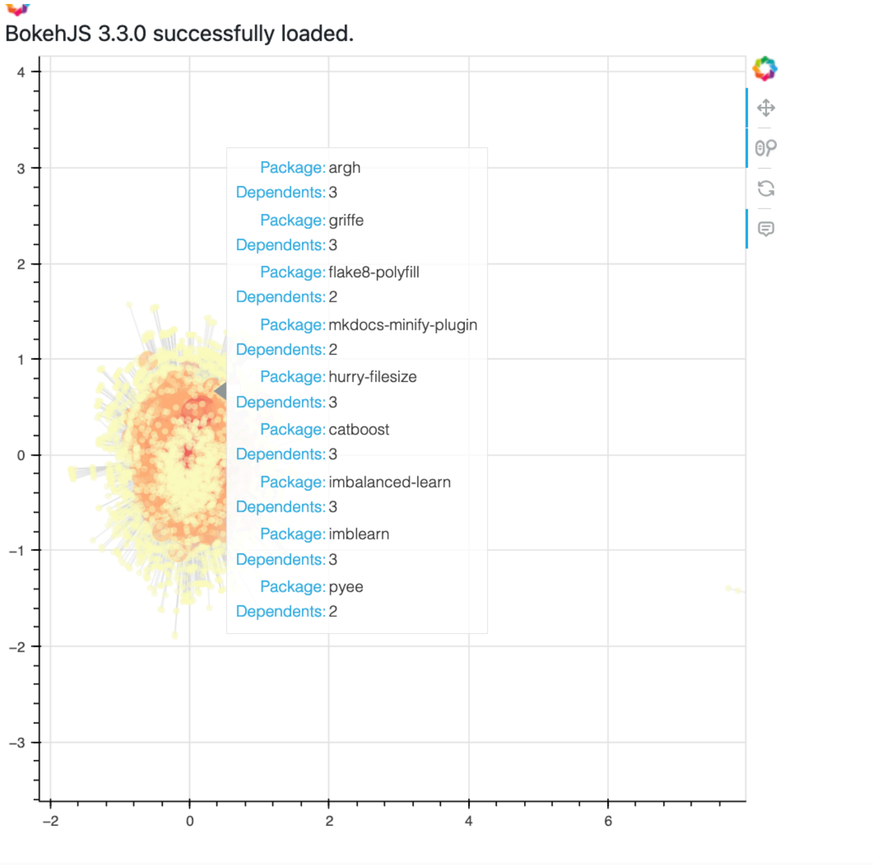
Software Developers

Producers

Who produces scientific software, and what role do they play in published research?

- Extract author lists from ~4000 software journal pubs (JoSS, SoftwareX)
- Extract developer profile from linked Github repositories
- Manually label ~3000 author -> developer pairs...
- Use DeBERTa (encoder) to train model predicting matches between developer and author... **We achieve an .97 f1** 👀

Science Software Dependencies



https://evamaxfield.github.io/rs-graph/viz.html

Most of this work is thanks to my talented students: Eva Brown, Isaac Slaughter and Lindsey Schwartz

Much of this work is funded by NSF award #2211275; with collaborators at URSSI (Dan Katz and Karthik Ram)

nmweber@uw.edu