

A photograph of a radio telescope array, likely the Murchison Widefield Array, set against a vibrant orange and yellow sunset sky. The foreground features a dirt road leading towards the horizon. Several large parabolic dish antennas are visible, with the largest one in the center-right foreground. The ground is covered in low-lying vegetation.

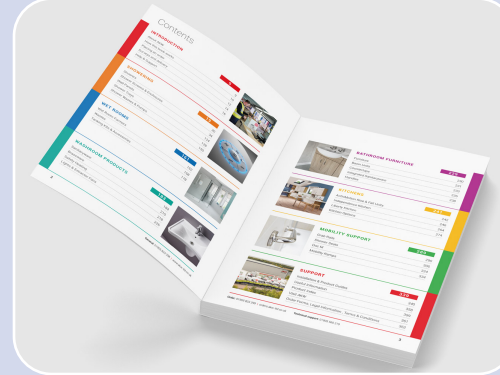
Overview of the Github Research Software Ecology

IRSER Community Meetup 15/16 January 2024

SSI Activity



Evidence
bank of
research
software
landscape
reports



Catalogue
of open
research
software



Case
studies of
research
software
projects

Research Objectives

A. Understand the Open Source Research Software ecology

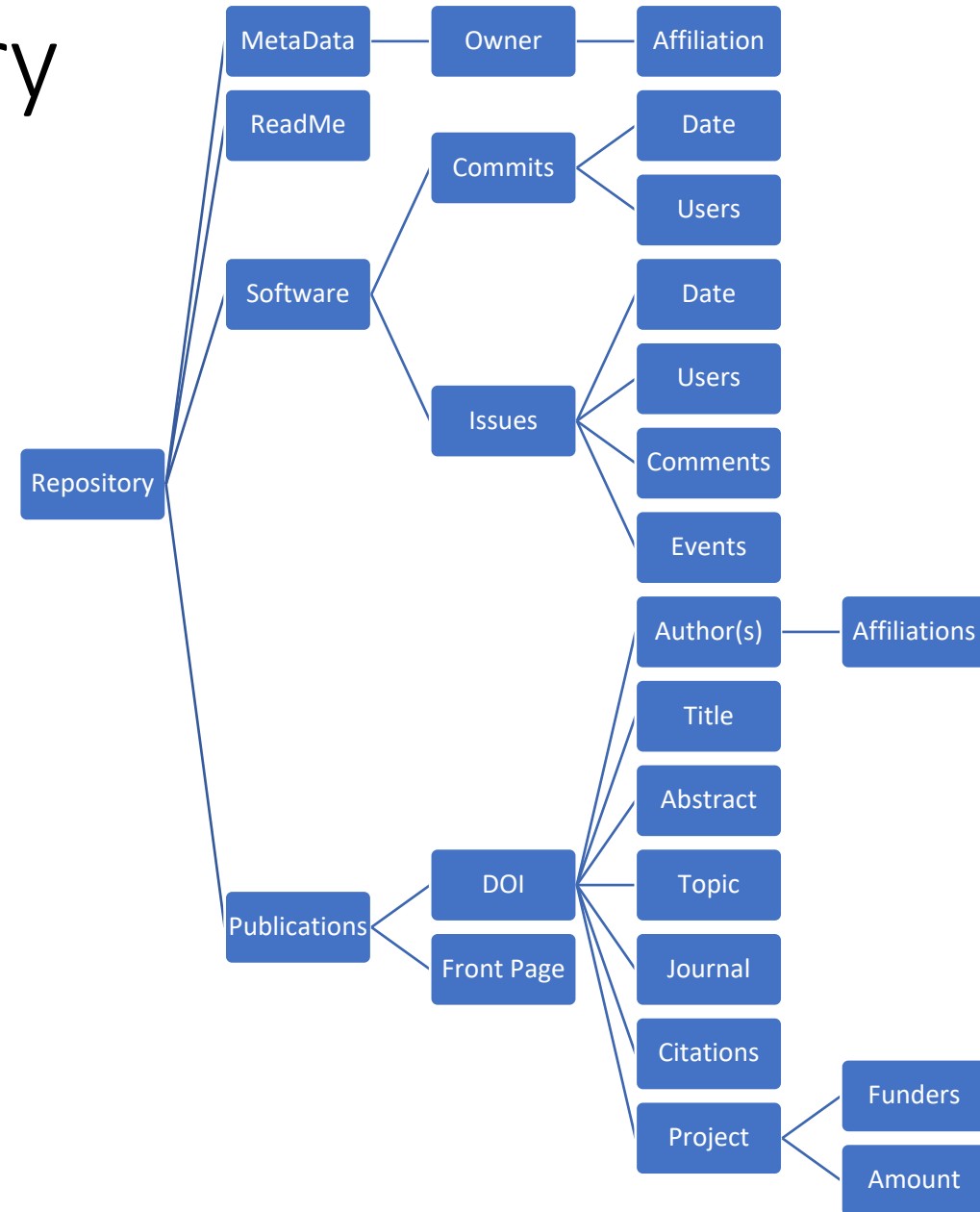
- What is the relationship between research, research software, research publications and research projects?
- How does software support or embody research methodologies, research methods, research impact
- What software development methodologies form common practice
- How do teams form and evolve around research software?
- How does (when does) software outlast specific funded projects
- How can we understand collaboration and contribution

B. Provide policy advice to funders and institutions

- How do researcher careers (teams) evolve with their research software contributions
- Effective ways to stimulate open science

Research Software Observatory Data Collection

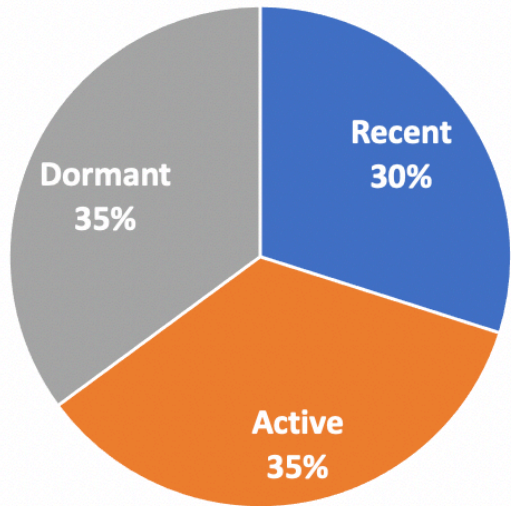
- Open Source Research Software is available in GitHub software repositories
- We can identify research-oriented software by querying GitHub for telltale **research signatures** regarding associated publications
 - DOIs, BiBTeX metadata, publication URLs
 - Also vv find Github references in PDFs



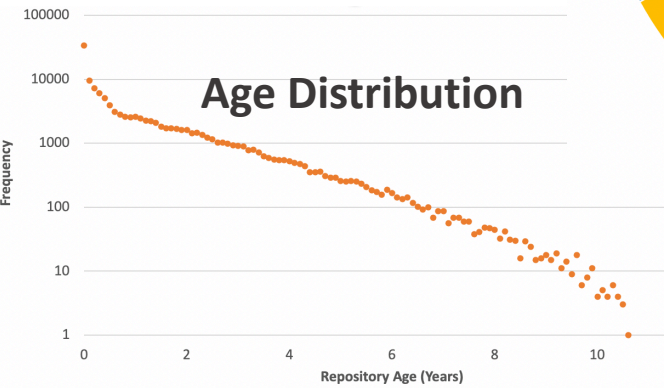
Collection Statistics

N=123893

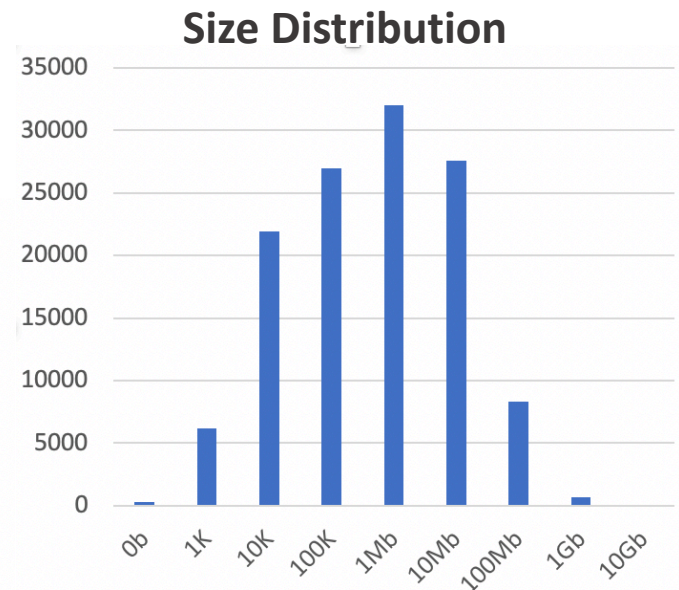
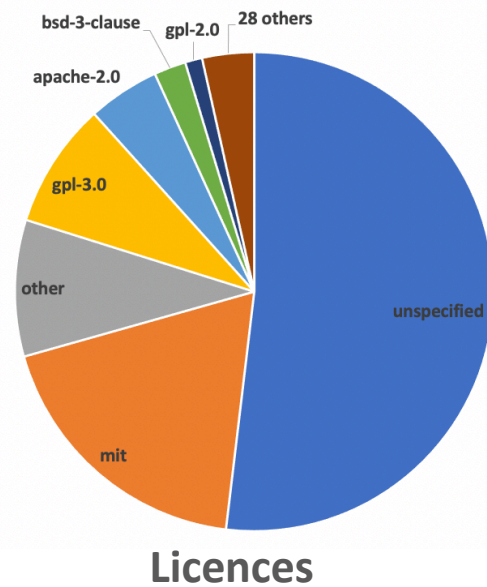
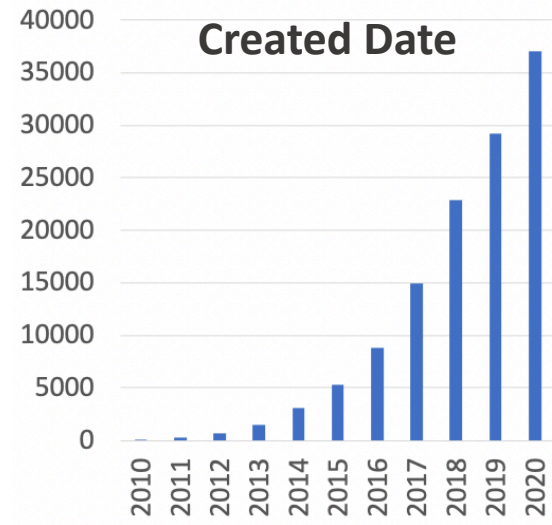
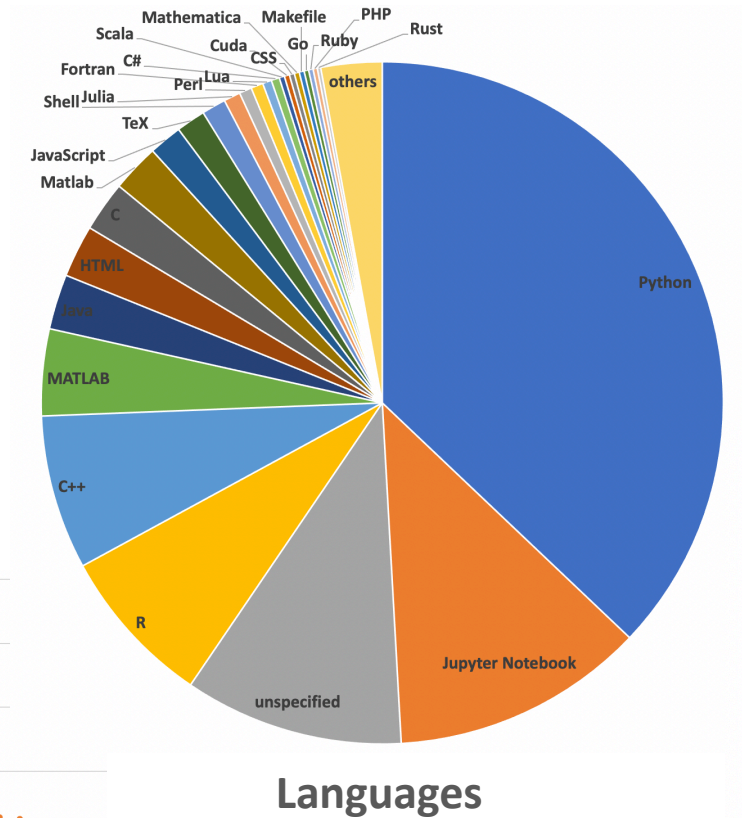
Census=2020



A repository is **active** if it has been updated in the last year. It is **recent** if it was created less than a year ago.



The repository age is the time between its initial creation and its last commit.

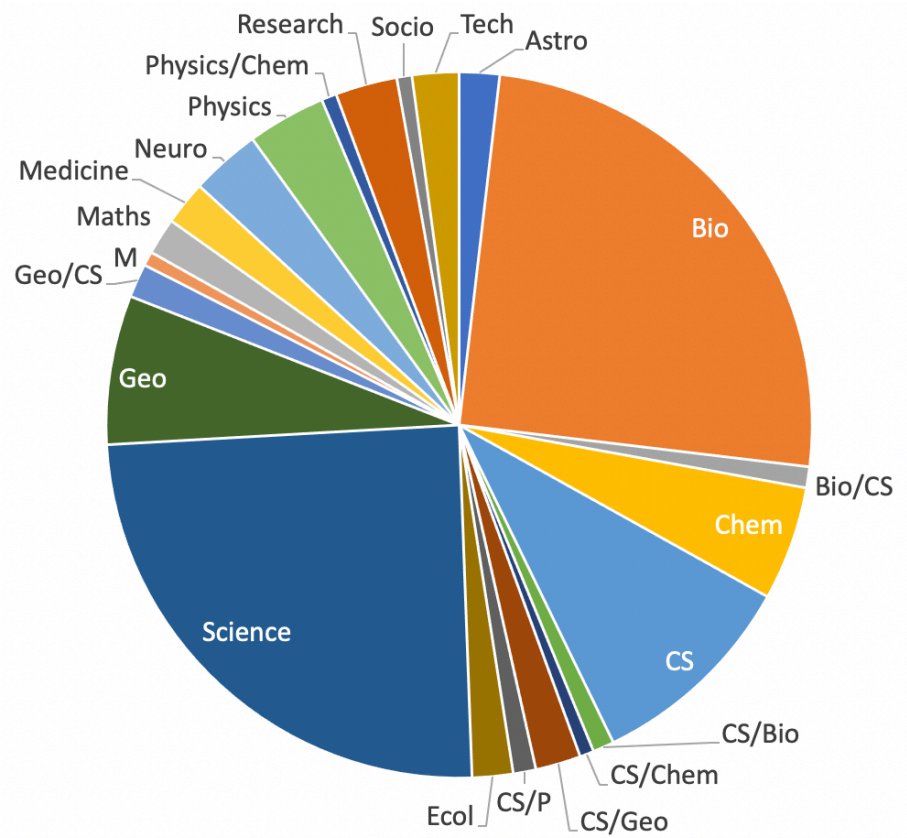


- 1-9Mb single user, 1 year, single paper
- 10-30Mb small team, 3 years
- 1Gb+ dominated by large data contents

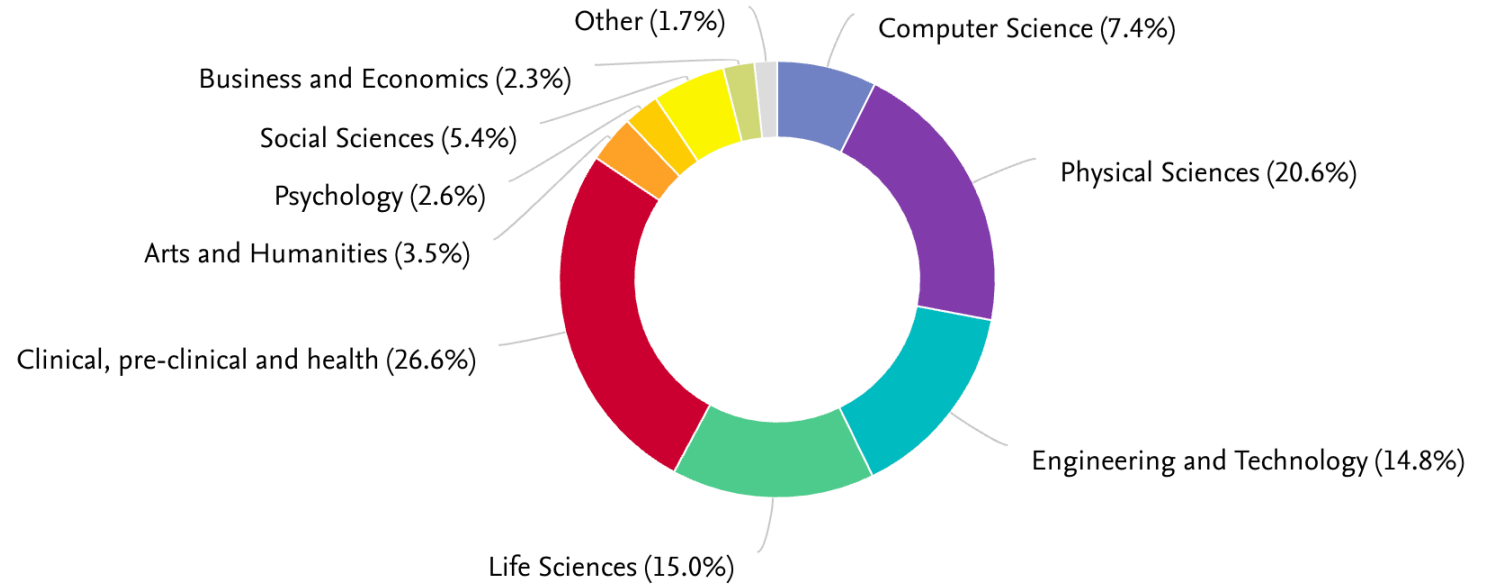
Discipline Alignment

The broad split in global research is in the proportion CS:Physical Science: Eng & Tech: Life Science: Health&Medical: Social & Human Sciences & Art :: 1:3:2:2:4:2

The rough split in Github research is currently unknown



Topics of Most Frequently Mentioned Publications (Freq>100, N=69)



Topics of All Publications THES World Ranking Top 100 (From Scopus, Same Period, N=5.9m)

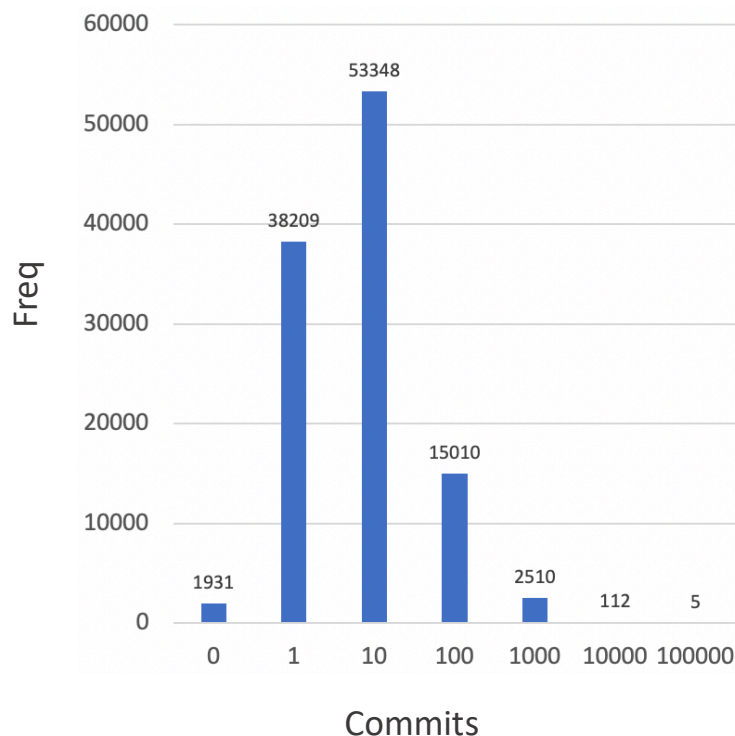
Activity Stats

380k accounts -> 83625 repository owners, 296510 other contributors. **30% of contributors** work on more than 1 repository.

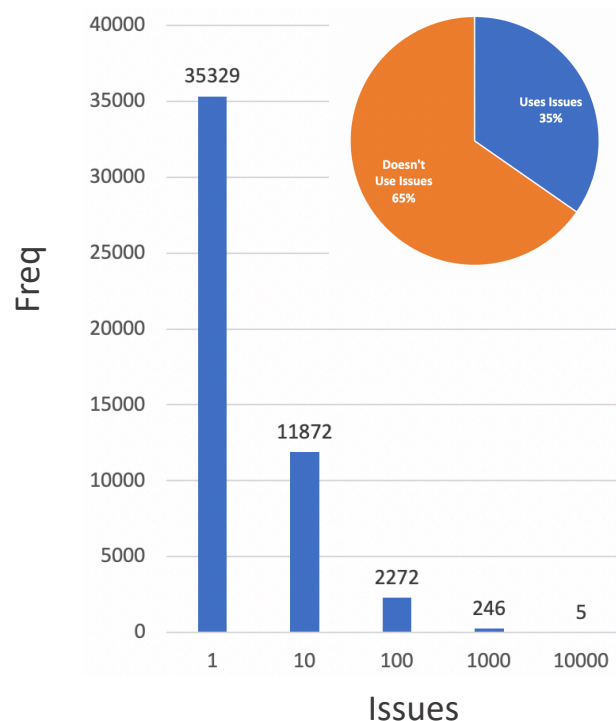
Most **common team size** is **1 or 2** – 90% of all repositories have 8 or fewer contributors.

In **75% of repositories** a single person performs almost all code commits, even though 71% of repositories have more than one team member.

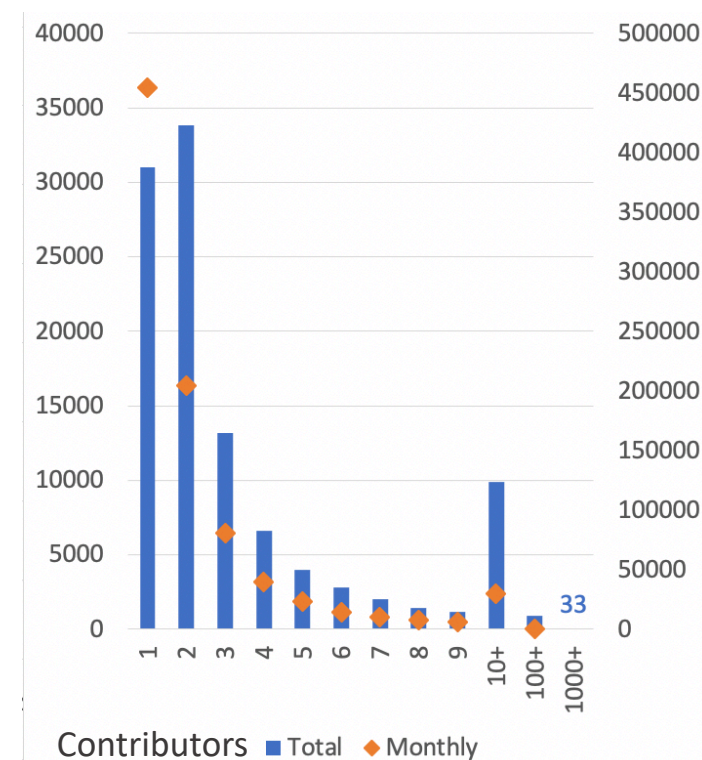
Commits Per Repository



Issues Per Repository



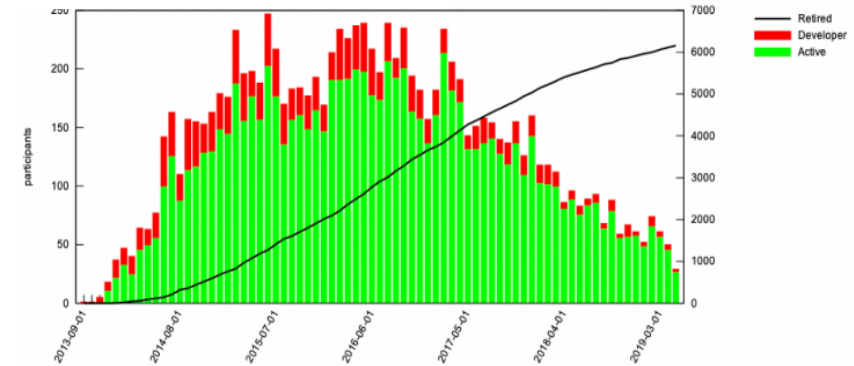
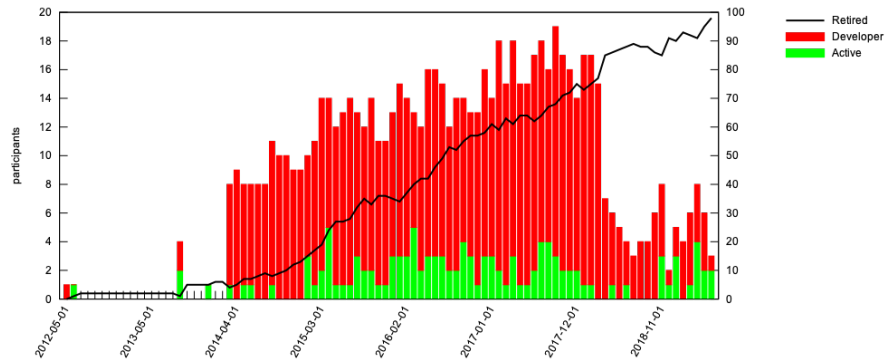
Contributors Per Repository



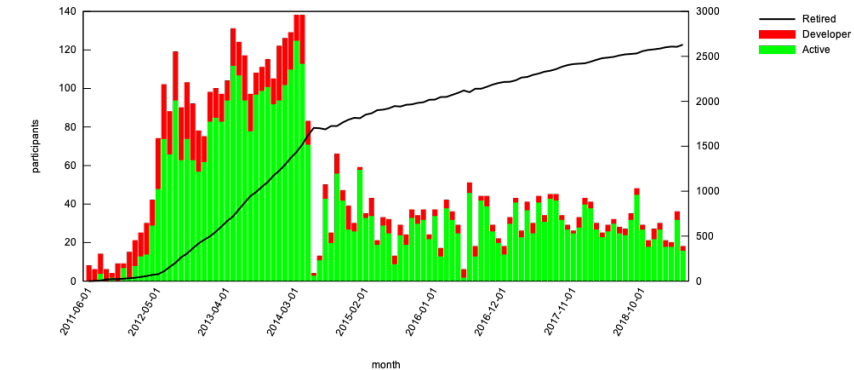
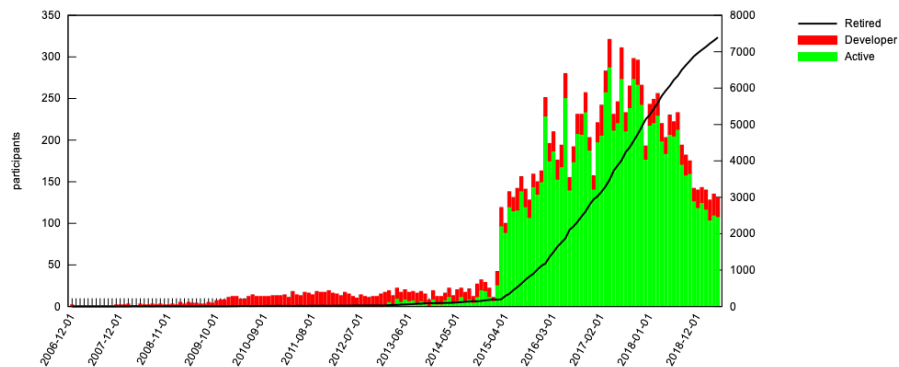
Average 0.5 commits/day. Most research repositories have 10-99 commits, **equivalent of 1 week to 2½ month's activity**

Contribution Roles

Github accounts that act as committers and issue closers are nominated “developers” (red) as opposed to those whose role is to raise and discuss issues (green).

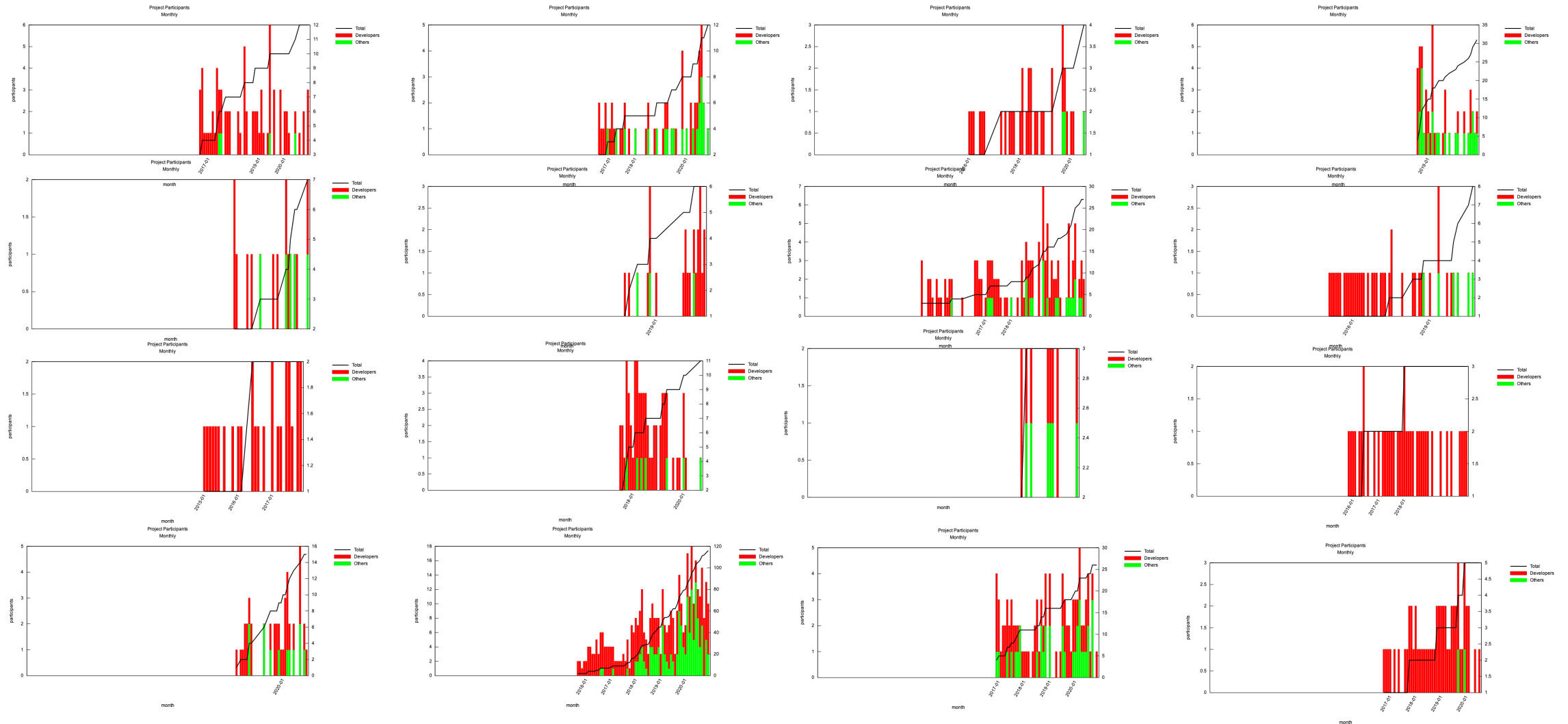


90% of repositories with <10 contributors are developer-dominated, but 89% of larger communities are developer-supported



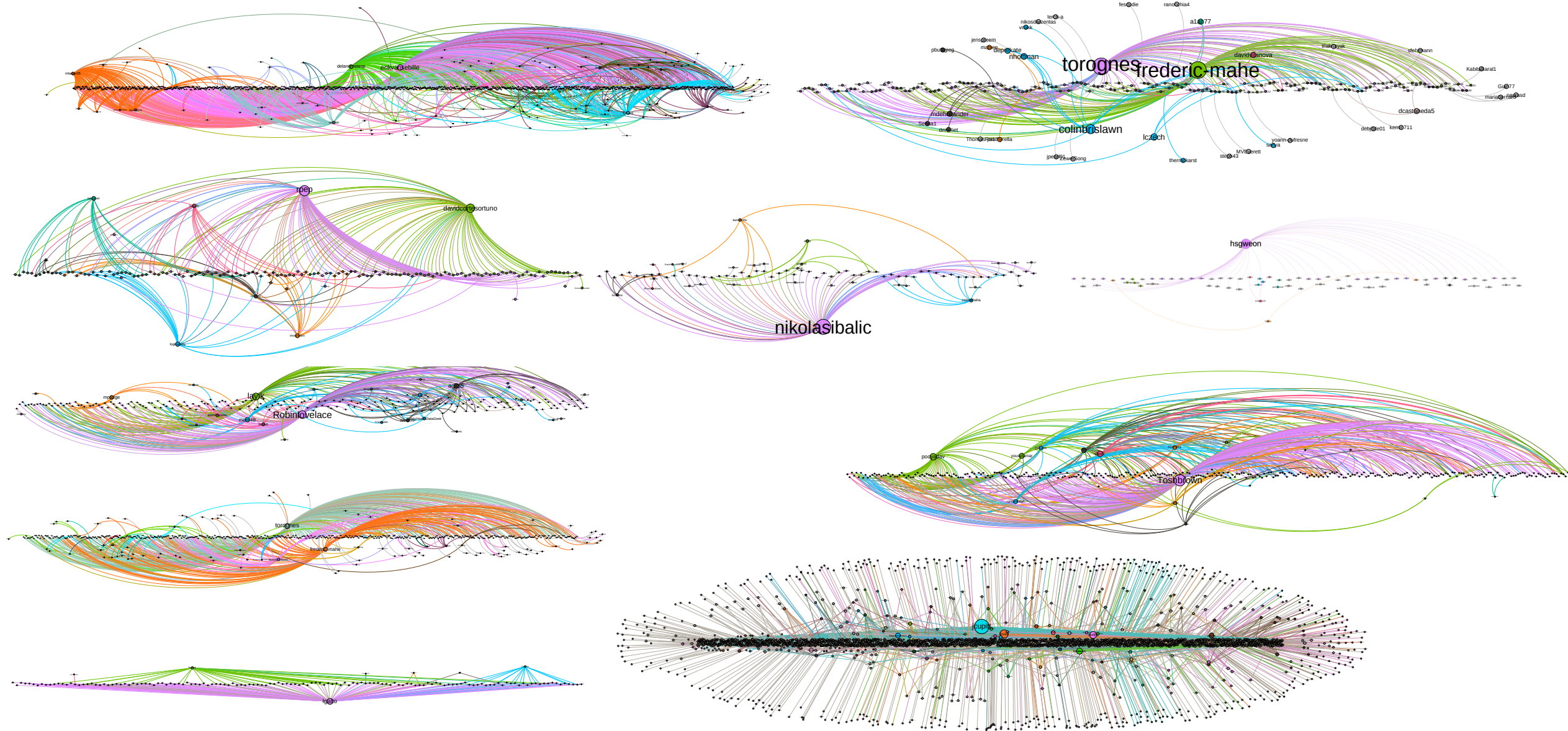
Communities may undergo a sudden phase change that causes them to gain or lose support

Sample Lifecycles of EPSRC Repository Communities



Much smaller communities; developer dominated; shorter lifespans

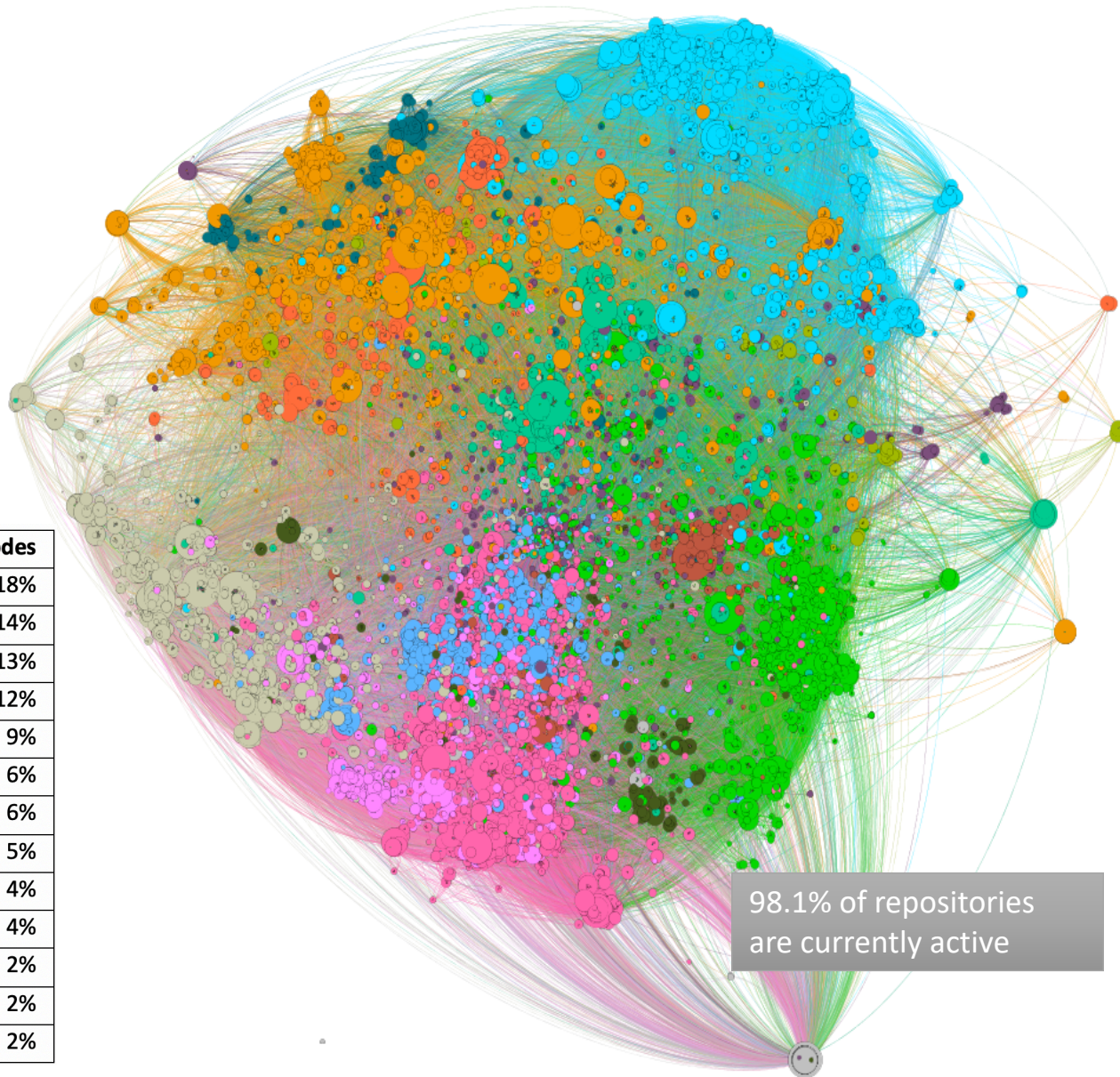
Community Interaction: Collaboration around Issues



Significant Reuse, Well Supported Open Science Software

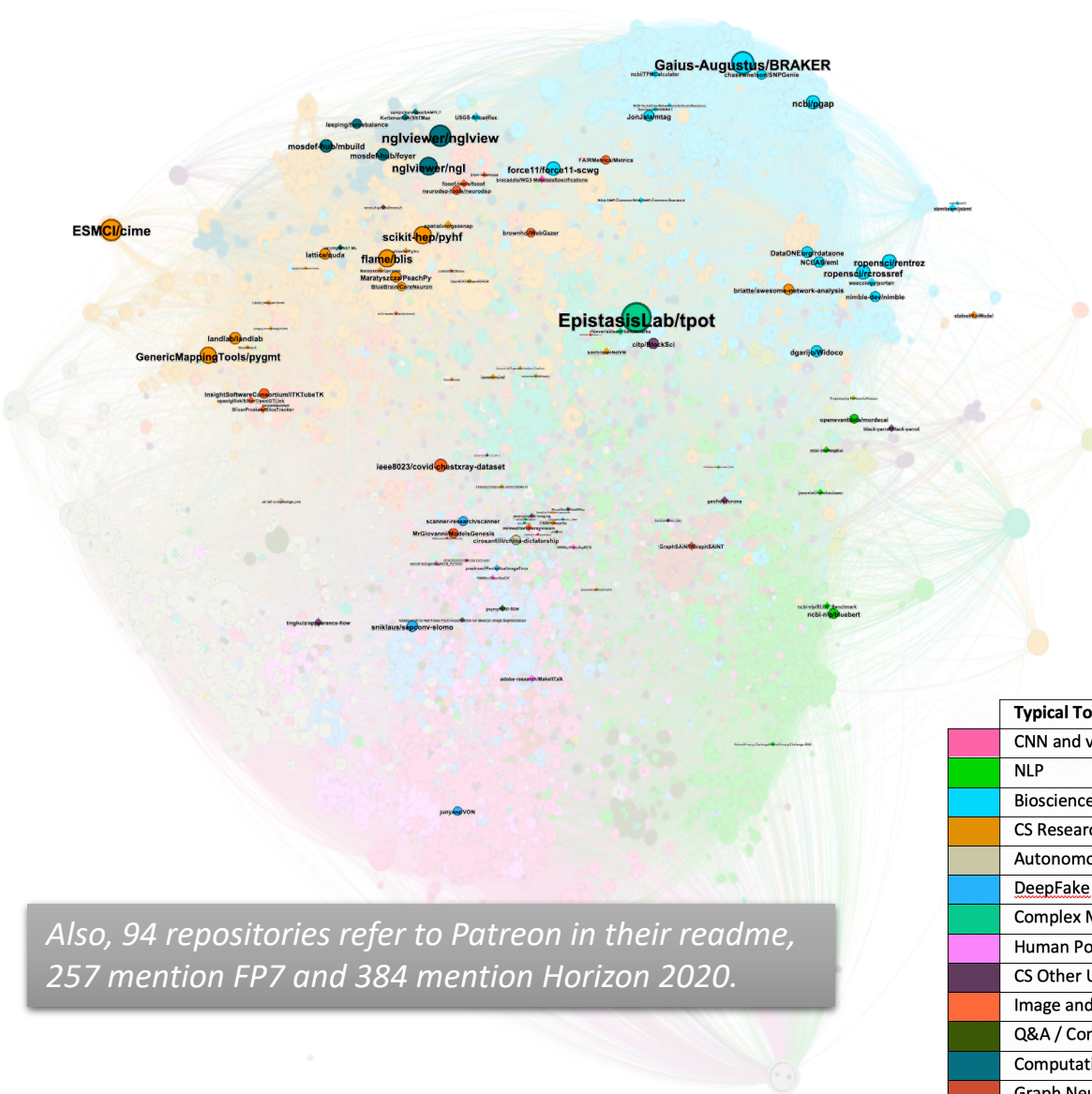
- There are 10k significantly reused research software repositories in GitHub
- Supported by 65k collaborators
- Network shaped by
 - research topic
 - software utility

	Typical Topic	Nodes
	CNN and variants	18%
	NLP	14%
	Biosciences / Data Structures	13%
	CS Research Utilities	12%
	Autonomous / Mapping / Odometry	9%
	DeepFake / GAN	6%
	Complex ML Infrastructure	6%
	Human Pose / Motion	5%
	CS Other Utilities	4%
	Image and Signal Analysis	4%
	Q&A / Conversation / Narrative	2%
	Computational Chemistry	2%
	Graph Neural Networks	2%



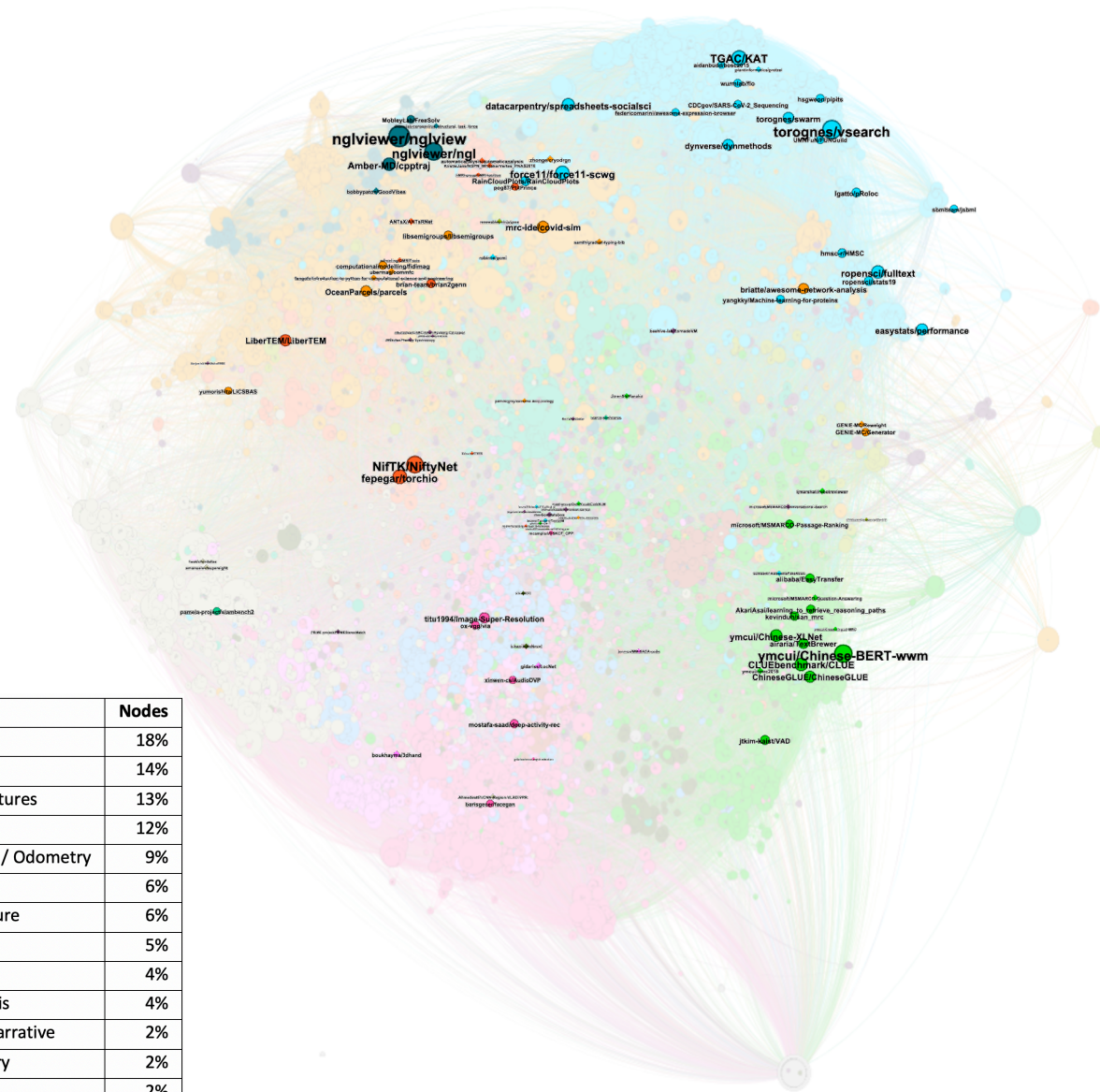
NSF/NIH Funded

acknowledges funder in README



Also, 94 repositories refer to Patreon in their readme, 257 mention FP7 and 384 mention Horizon 2020.

UKRI Funded



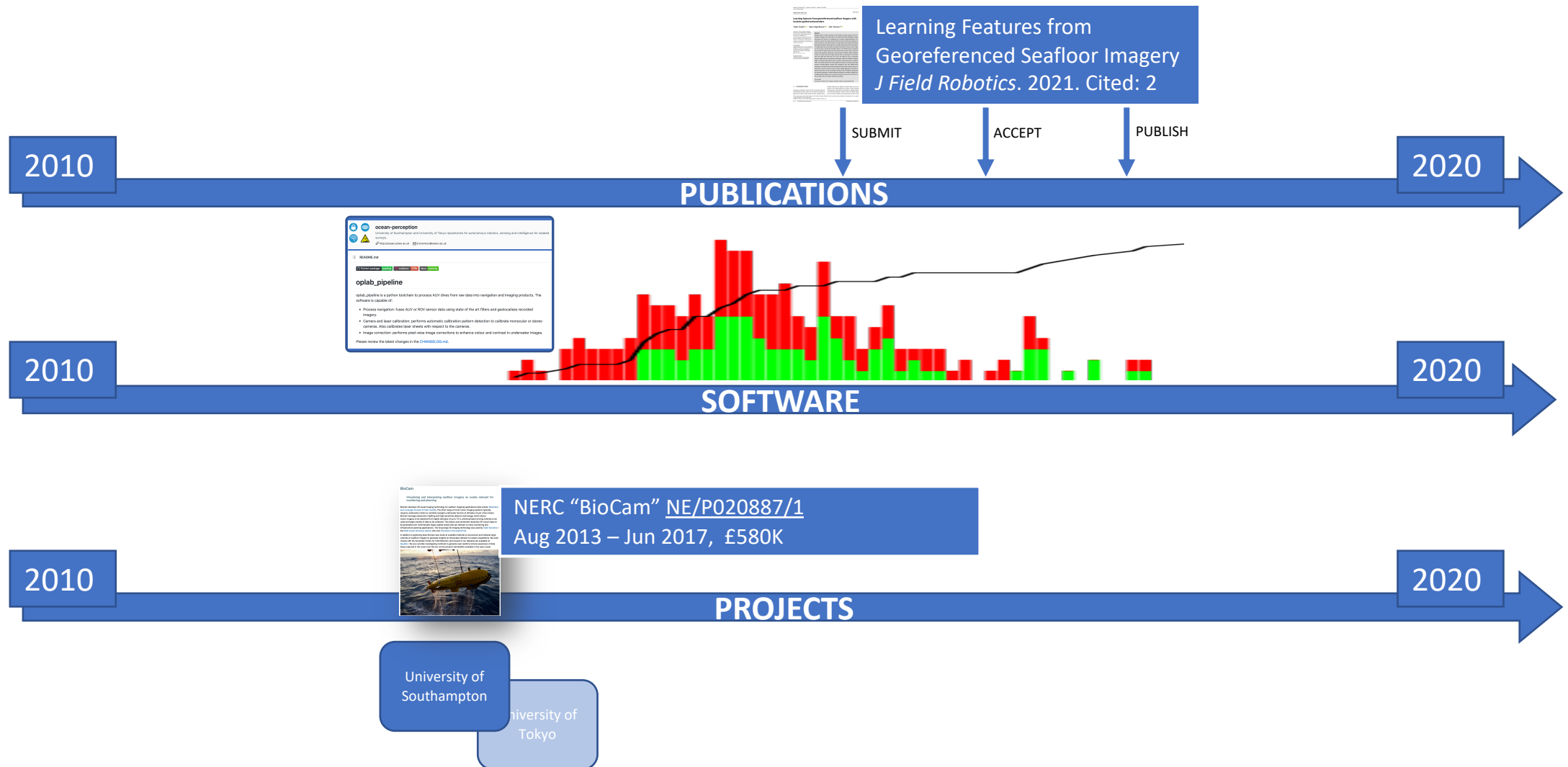
	Typical Topic	Nodes
	CNN and variants	18%
	NLP	14%
	Biosciences / Data Structures	13%
	CS Research Utilities	12%
	Autonomous / Mapping / Odometry	9%
	<u>DeepFake</u> / GAN	6%
	Complex ML Infrastructure	6%
	Human Pose / Motion	5%
	CS Other Utilities	4%
	Image and Signal Analysis	4%
	Q&A / Conversation / Narrative	2%
	Computational Chemistry	2%
	Graph Neural Networks	2%



Software Biographies

- Describing software is used and developed in the research ecology
- Describing how developers function as active researchers, authors and project investigators
- Describing how active researchers function as developers
- Describing how funded projects benefit from software

Diagramming A Software Bio



Research Software Biography Patterns

as seen in the Github Repository

- A. encapsulates the “working data analysis” (*i.e.* research method) of a piece of research
 - Subsequently tidied up for FAIR access as part of the publication process
- B. an independent package of research analysis functionality generalised from one or more earlier research activities / papers / repositories (A)
- C. implementation of a novel software product that is used as the focus of research (*i.e.* not the research method itself)

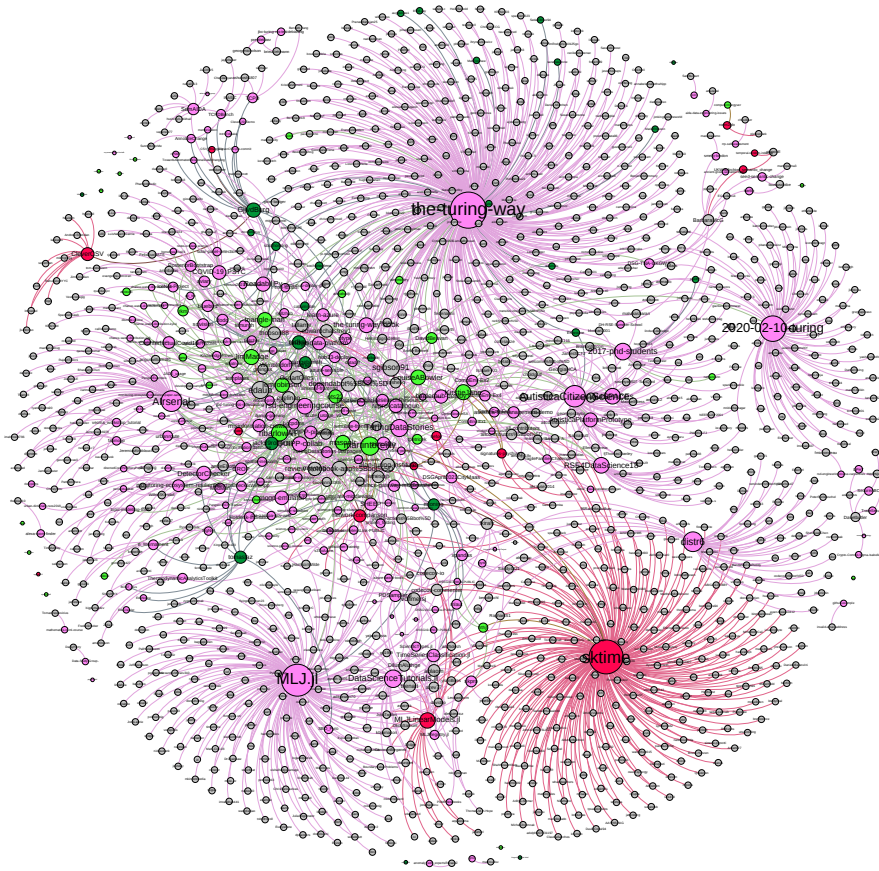
Sustainable UKRI funded software

- Examine the identifiably UKRI-funded research repositories with the biggest associated contributor communities using a “software biography” method
 - What was the relationship of the funding to the software
 - What was the relationship of the software to the research publications
 - What role did the project PIs and CO-Is play in the software development
 - What role did the software developers play in the research project and its publications
- What has led to large, sustainable science s/w collaborations?
 - a) contribute to an independent software collaboration
 - i. allow your students and postdocs to work on already successful 3rd party software
 - ii. be an early adopter of 3rd party software and write up research papers with the software authors
 - iii. create and promote your own software with your own team and attract your own developer community
 - b) Work with an open science community such as rOpenSci
 - i. develop your own research software and subsequently donate it to an "open science community" who help you to package it up for the world
 - ii. create your own significant collaboration with an explicit "open science" goals and generate useful software as an outward looking part of your research process e.g. Earlham Institute, Grantham institute, Turing Institute, rOpenSci

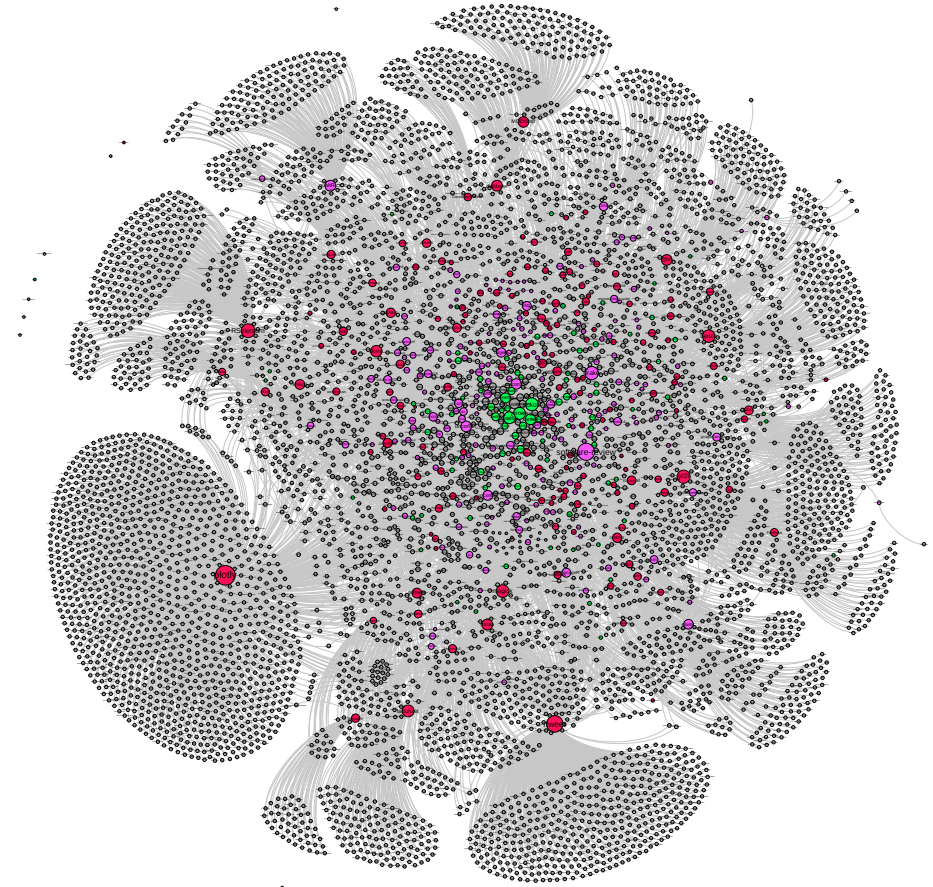


Funded Open Science Collaboration Networks

research / other repositories (red vs pink)
official / other contributors (green vs grey)



Alan Turing – 230 repositories, 1245 contributors

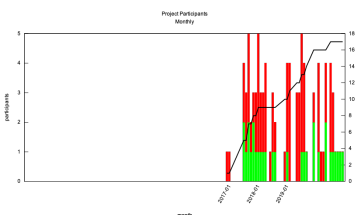


rOpenSci – 358 repositories, 6738 contributors

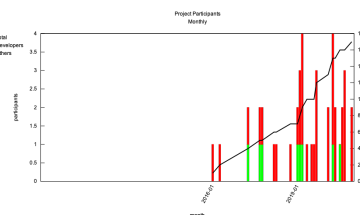
Research Beyond the Academy

Innovation & Startups

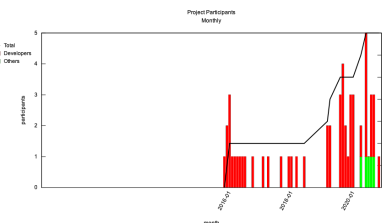
- Some high profile startups on GitHub
 - Only some repositories on GitHub
 - Only some repositories open on GitHub
- Teams may be small
 - It's not just public research being worse than the private sector!!
- Software may be short duration or project-specific utility
 - It's not all sustainable hero-software
 - How does research software team compare to a startup innovation team?
 - How does project funding compare to startup funding rounds?
 - How does tool use, training and career mobility compare?



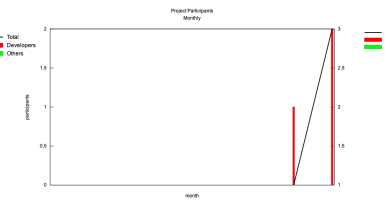
gocardless/draupnir



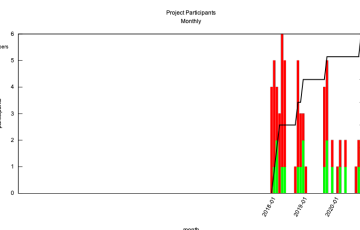
gocardless/homebrew-taps



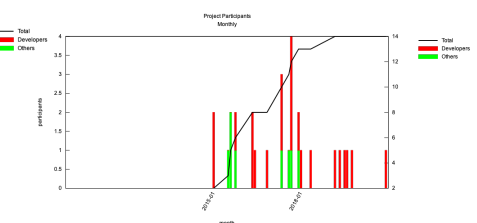
redsift/sandbox-javascript



riverlane/deltalanguage



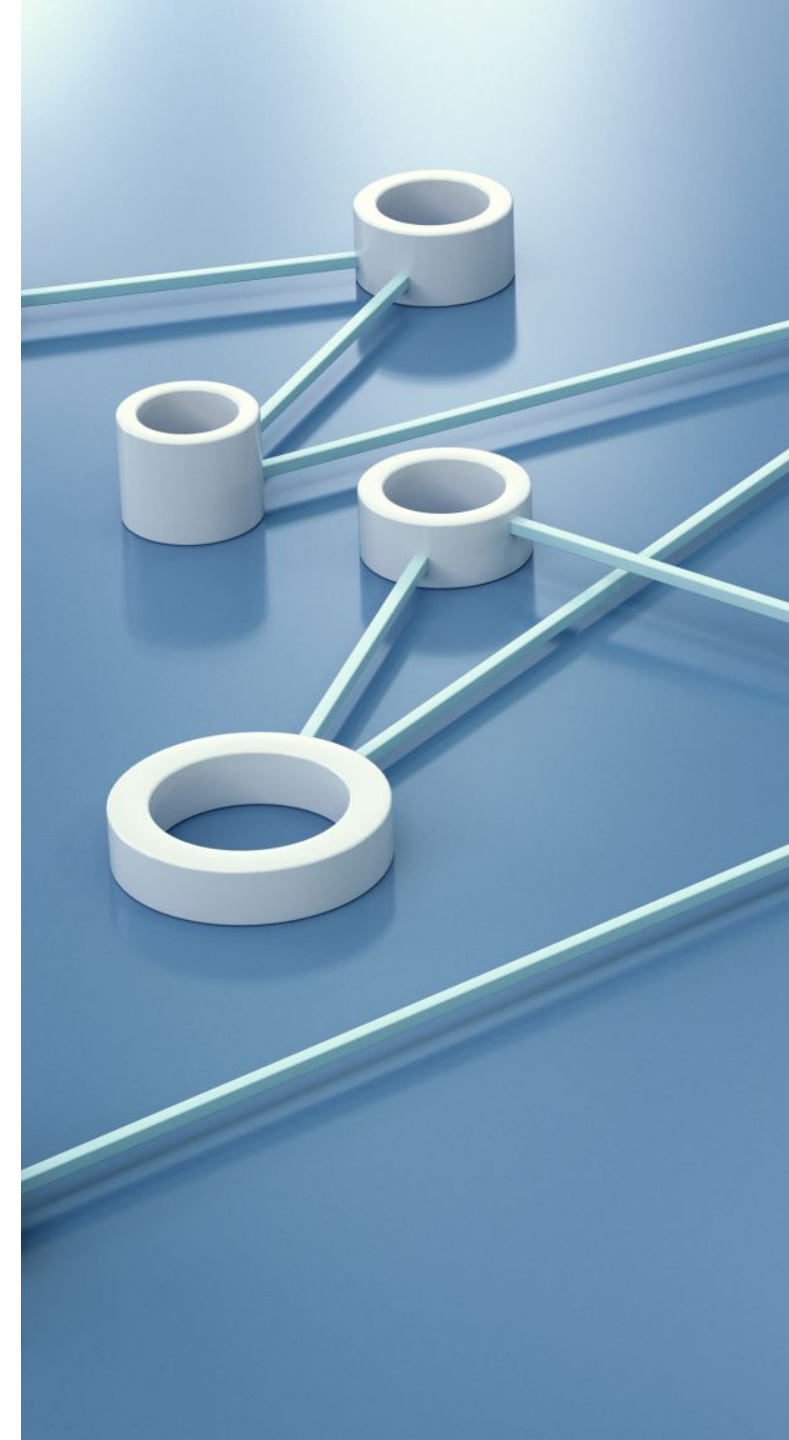
thread/routemaster



what3words/w3w-python-wrapper

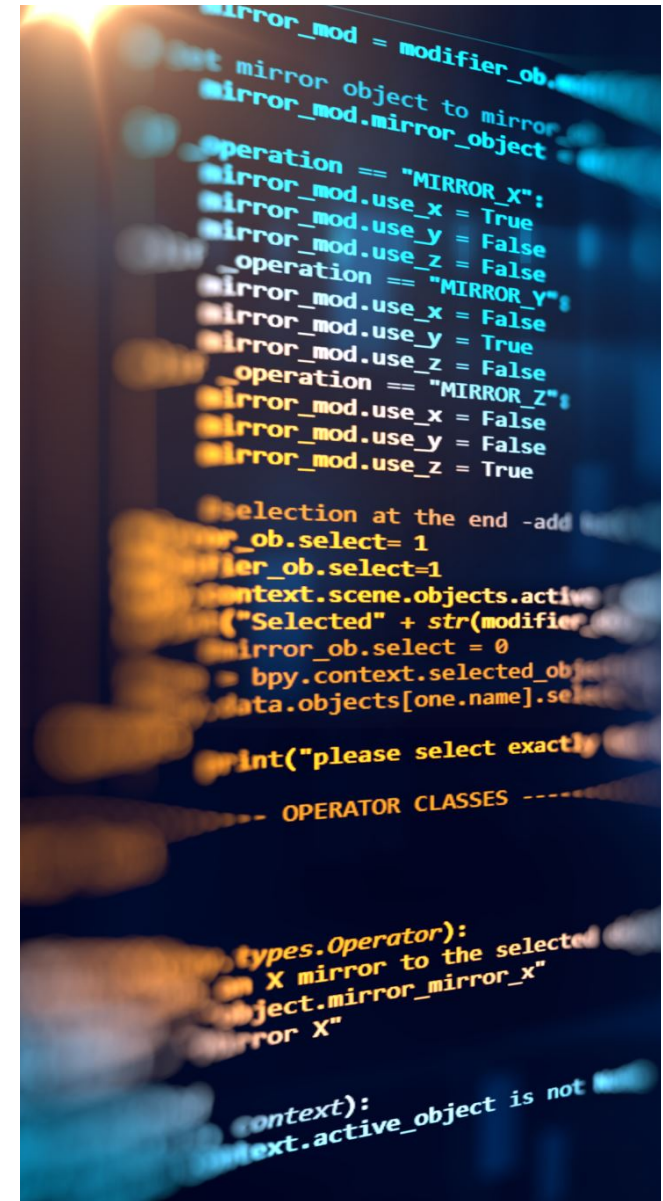
Large Language Model Activity

- Since the launch of ChatGPT in November 2022, there has been a huge amount of activity around Large Language Models
- The big tech AI companies (Google, Microsoft, Meta, X, OpenAI and to some extent Apple) have all been developing their own proprietary Language Model products and releasing them in their own commercial offerings (e.g. Bard, Grok, Microsoft Office Co-pilot).
- However a huge amount of third party activity has emerged around the (nominally) open OpenAI ChatGPT API and also around the (actually) open source LLaMA and ChatGLM foundation models.
 - This activity is visible on Github



Methodology

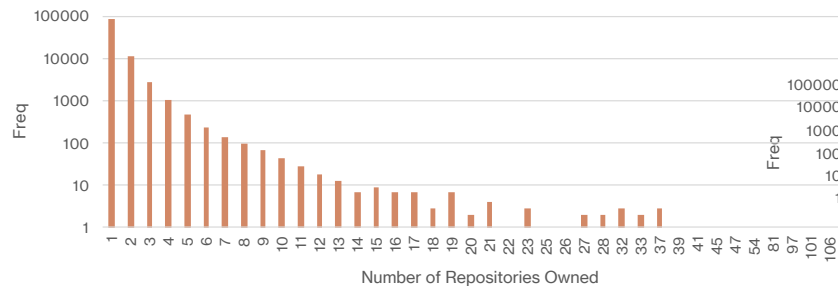
- Search Github for all repositories that match “LLM” or “GPT” since Nov 2022
- For all candidate repositories (there are false positives *e.g.* repositories about boot software for GUID Partition Table boot software) extract
 - Software Bill of Materials to see the dependencies of each project
 - Basic repository descriptive statistics (*e.g.* description, owner, number of commits, number of forks)
 - README.md *i.e.* the repository home page, giving a fuller description
 - List of all commits to determine significant contributor community



Initial Results

- 127,142 candidate repositories
 - Big boost after release of ChatGPT API
 - 11.5k appear to be research-oriented (DOI / arxiv / inproceedings)
- 85 repos have >1000 forks
 - 638 have >100 forks
- 1712 repos have >1000 commits
- 201921 different package dependencies listed in SBOM
 - 78 NLP packages used >10x
- The clustering of high-contribution developers is quite pronounced.

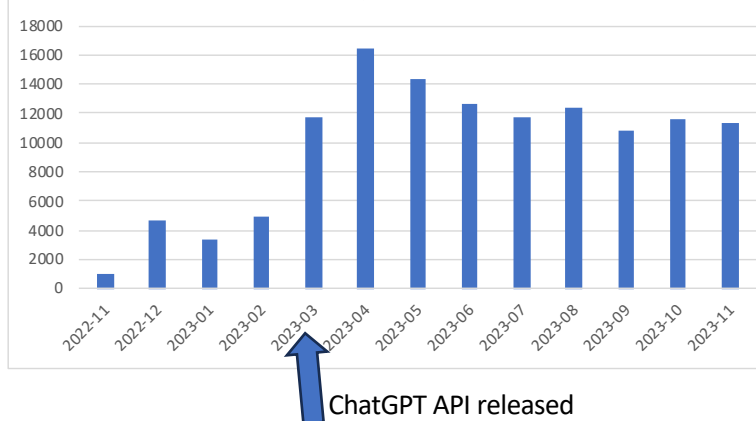
Account Owns 'N' LLM/GPT Repositories



Number of Forks of Each LLM/GPT Repository



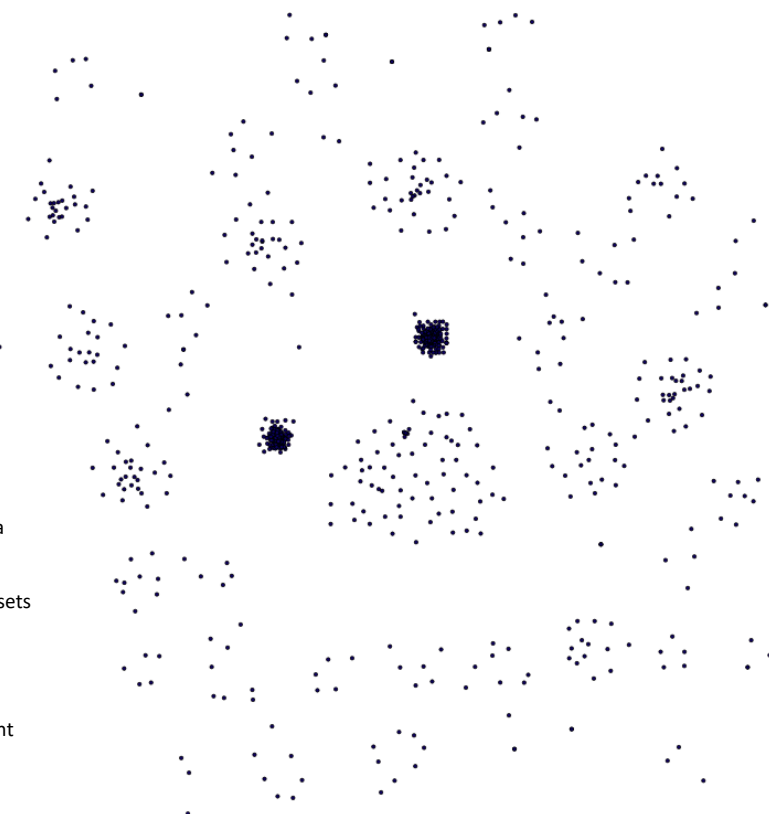
Number of LLM/GPT Repositories Created Per Month



40900	openai	1471	tensorboard	220	torchtext
14966	langchain	1448	tensorflow-text	212	optimum
10736	torch	1182	gpt-3-encoder	211	openai-gpt3-java
6120	chatgpt-tool-hub	1376	llamahub	204	litellm
5942	chatgpt	1133	tensorflow	198	auto-gptq
3650	datasets	1103	openai-edge	189	tensorflow-datasets
3380	accelerate	996	peft	183	async-openai
3271	safetensors	809	diffie-hellman	181	edgegpt
3177	pinecone	705	gpt3-tokenizer	172	ruby-openai
2546	revchatgpt	695	go-openai	123	llm
2307	gradio	585	gradio-client	118	openai-php/client
2266	openaiauth	554	pytorch-lightning		
1685	bitsandbytes	493	gpt4all		
1673	openai-whisper	347	torchmetrics		
1584	gptcache	343	chatgpt-next-web		
1523	tensorflow-hub	339	loralib		

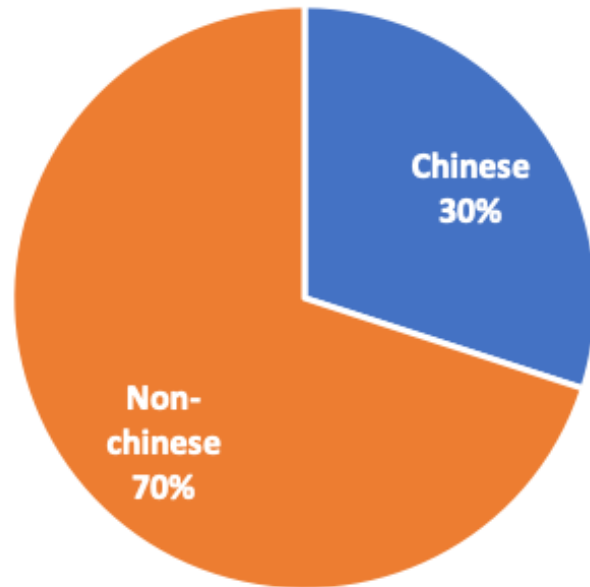
Main AI package dependencies

Collaboration of high contribution developers

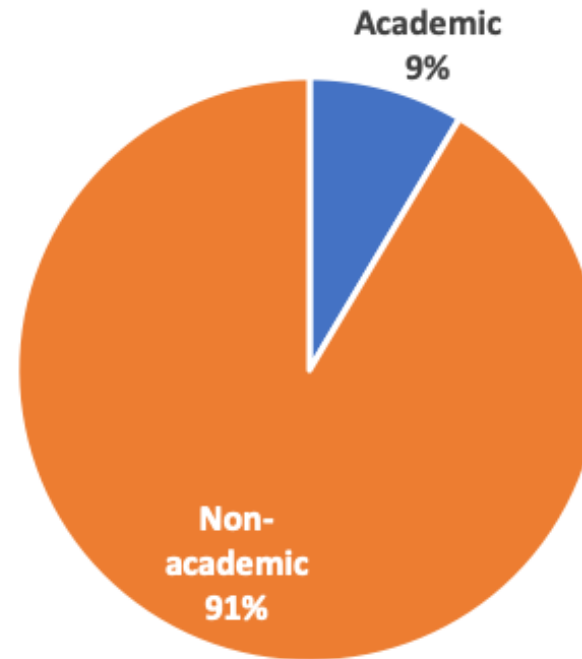


Background

Language of Readme



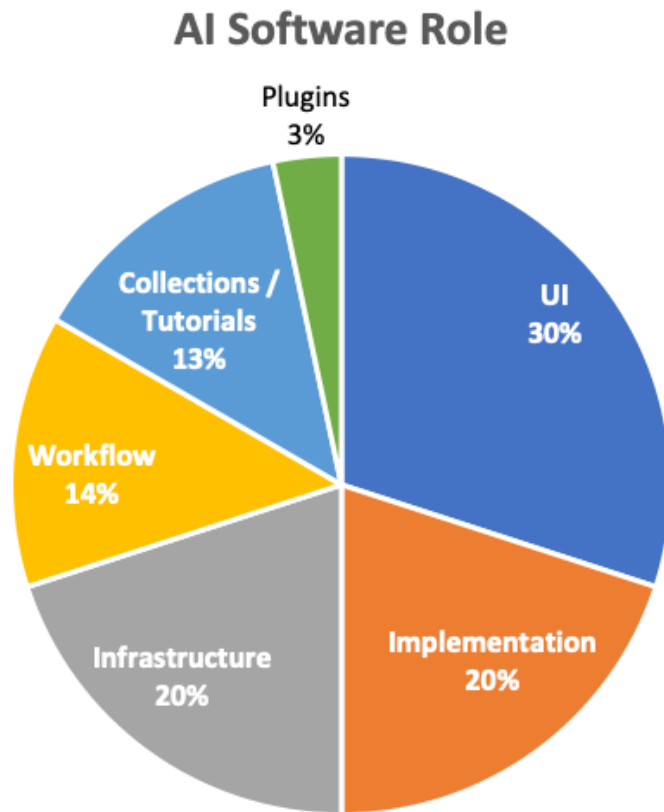
Source of Repo



- Chinese repository = 2 or more lines in Han script in README.md
- Academic repository = README match on doi:, arxiv, @article or @inproceedings, .edu or .ac.uk

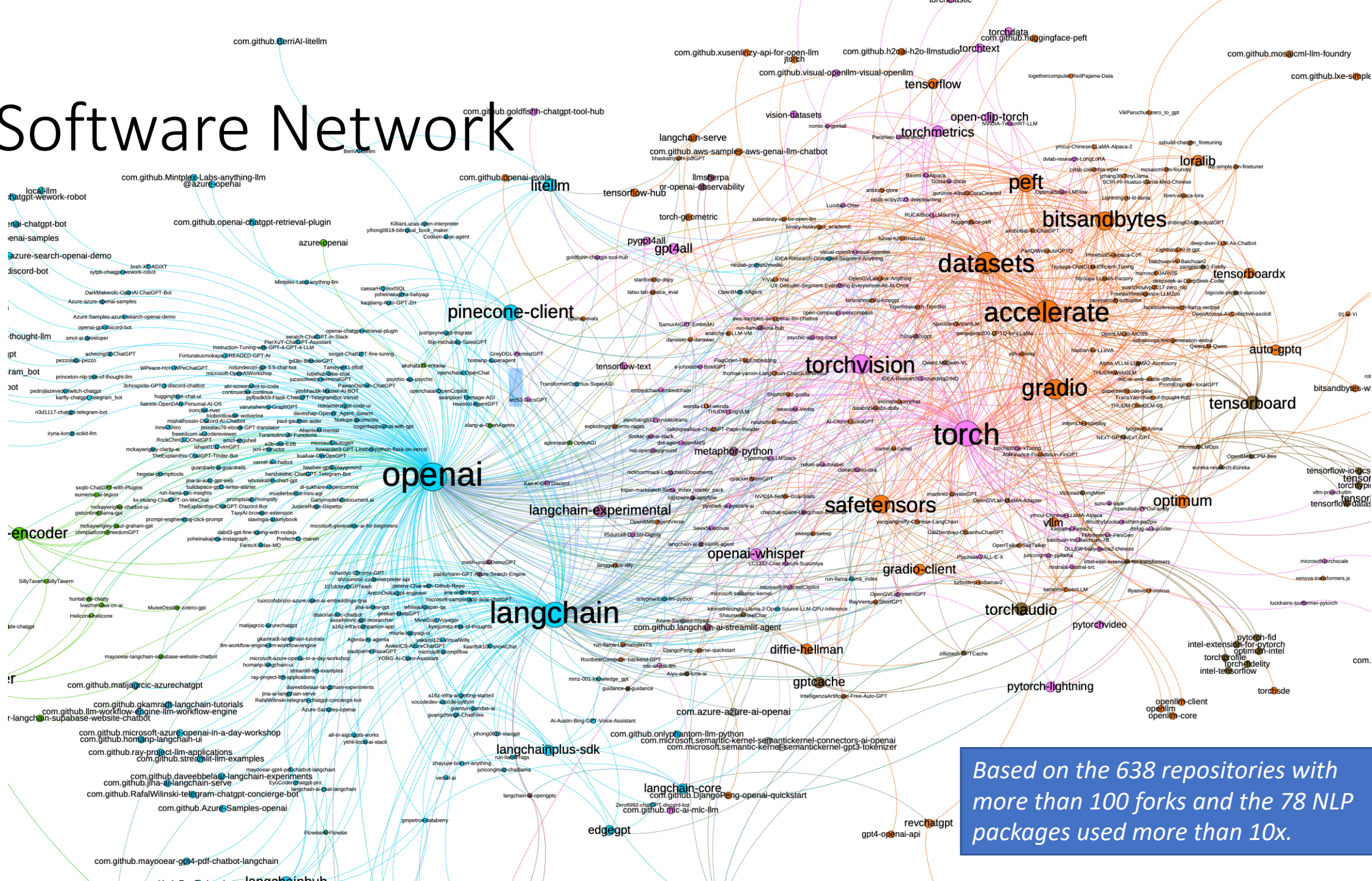
Most Popular* Repos

*currently based on top 30



- UI – a useful app using the ChatGPT API
- Implementation – code that implements a transformer-based language package *e.g.* Llama2 or ChatGLM or a derivative
- Infrastructure – code that allows you to assemble the components of an LLM package into a new AI product
- Workflow – code that allows you to deploy multiple LLM stages / activities into a coherent production stream

AI Software Network



Based on the 638 repositories with more than 100 forks and the 78 NLP packages used more than 10x.