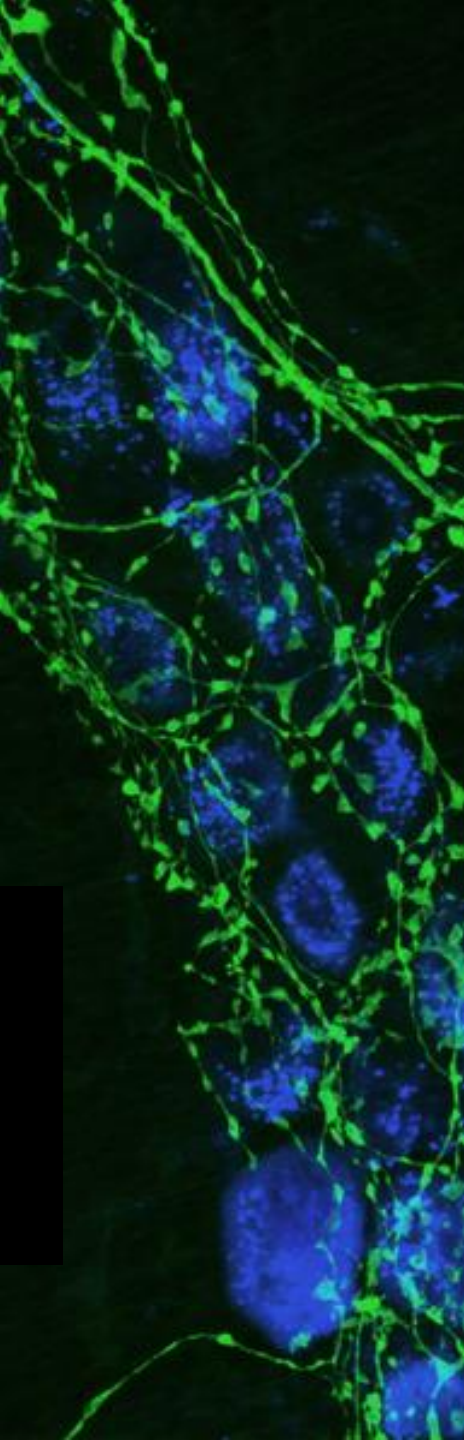
OPEN SOURCE IN SCIENCE

How contributability, transparency, and other open source practices can revolutionize scientific research



@lisancao

LISA N. CAO
SIMON FRASER UNIVERSITY

A GENTLE INTRODUCTION TO OPEN SOURCE

Open Source Software

- **Publicly available source code**
  - Within the license permissions, be able to be reviewed, reused, and 'remixed'
    - Licenses must not require royalties
  - Can be contributed to by community members of different expertise from all over the world
  - Has forums where discussion such as new ideas, vulnerabilities, and solutions are discussed
  - Ideally, comes with good documentation
  - Can be forked
- **Free Redistribution**
- **Technology-Neutral**
- **Has options to protect the author's source code depending on license**
- **Less emphasis on the individual, and more on the community**

CHAOSS

FOSS Project

free and open source software project

open source initiative

@lisancao

A GENTLE INTRODUCTION TO RESEARCH

... as an industry

Scientific Research

- **Must be:**
  - Reproducible
  - Impactful/publication worthy (not really)
  - Involve experimental design, meet research standards, etc
  - Statistically analysed

- **Should be:**
  - Assessible by third parties (peer reviewed) from the bottom up (open data)
  - Accessible to the public, or at least other researchers (open access)
  - Driven by innovation, sound scientific practice, technological advancements, and collaboration
  - **RE-PRO-DU-CI-BLE!!**

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

open science

OPEN ACCESS

R consortium

@lisancao

THE OPEN SCIENCE FRAMEWORK

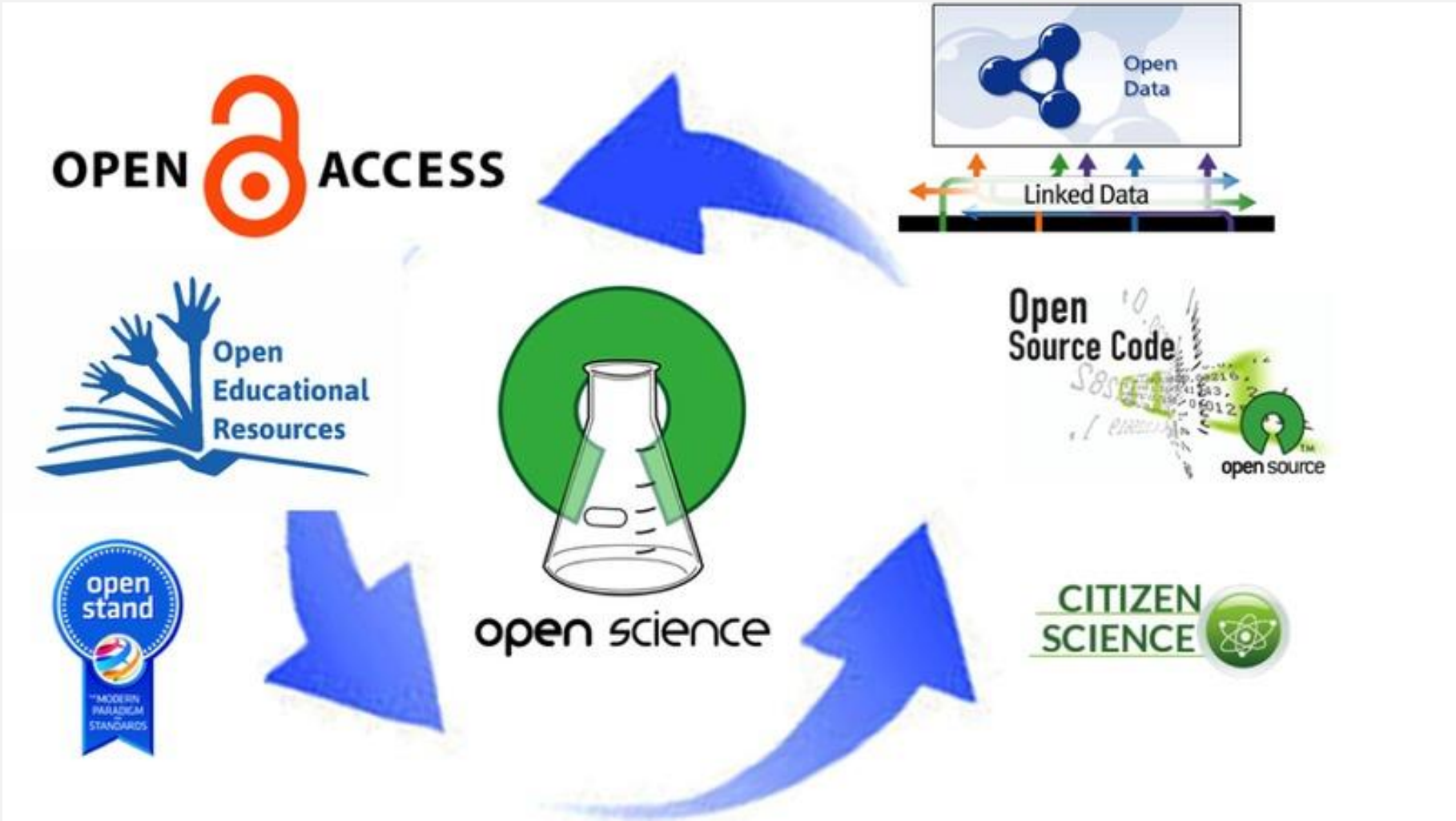
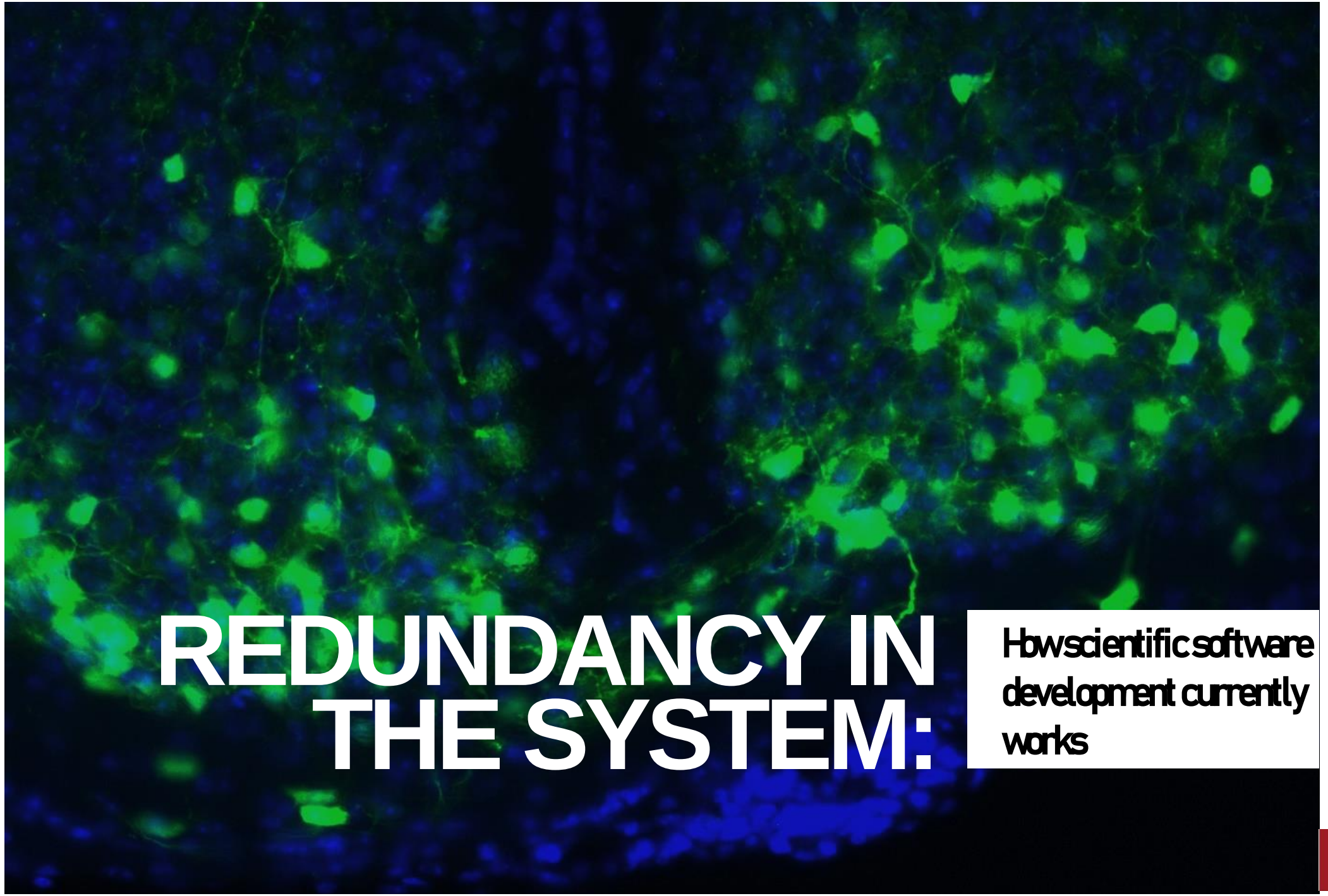


Image Source: http://www.sci-gaia.eu/osp-enab/

- Open Science
  - Open Educational Resources
  - Open Access
  - Open Peer Review
  - Open Methodology
  - Open Source
  - Open Data

Image Source: Andreas E. Neuhold

@lisancao



REDUNDANCY IN THE SYSTEM:

How scientific software development currently works

@lisancao

CODING WHEN YOU'RE NOT A SOFTWARE DEV

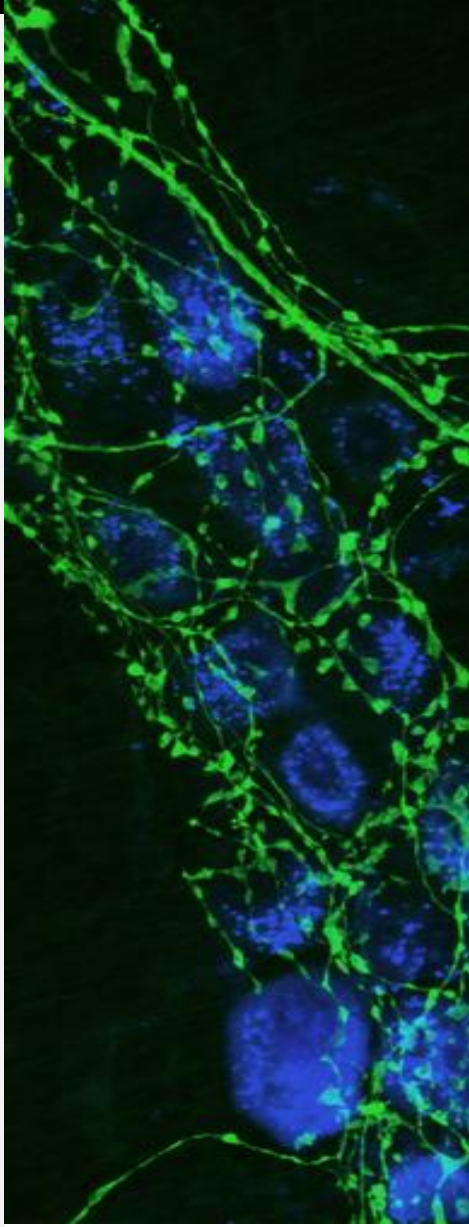
- My data has been collected, but I have issues with analysing it due to it's unique nature
  - E.g. a special imaging filter for x brain area, or cell sorting based on these principles, or time series analysis

- In order to solve this problem, I'm going to write some code in x program

- (Performs the analysis)

- (Publishes the results in a peer-reviewed paper that doesn't go through my code line-by-line)

- (Leaves the code on local server for either future use or abandons it entirely)
  - OR: "This is really good I'm going to attempt to license and sell it"



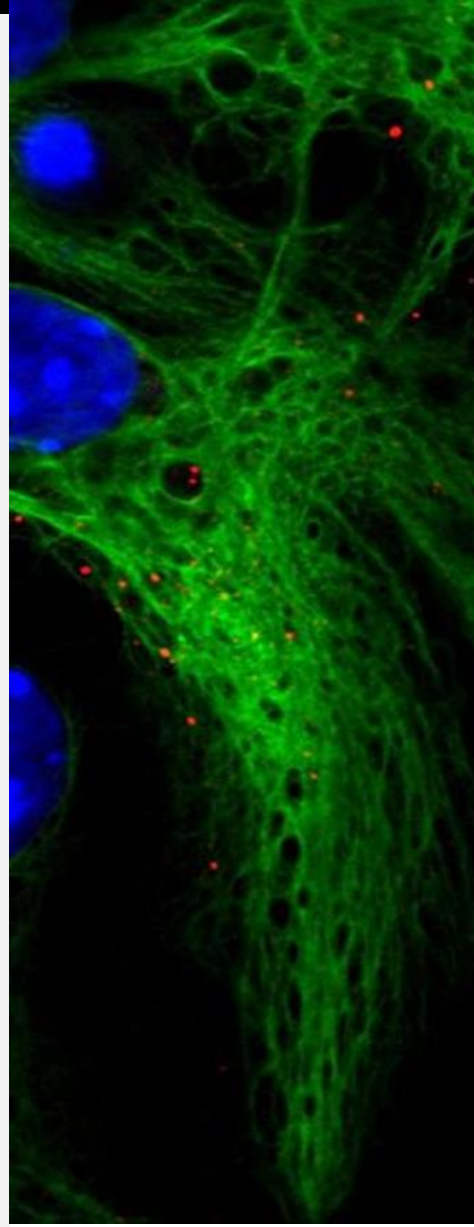
CODE REDUNDANCY IN SCIENCE

Within niche fields, there are unique software needs that are only understood by researchers in that particular field. Therefore..

- **You get a lot of code written by informal programmers**
  - Code which may not be the most efficient

- **You get several labs presenting the similar solutions, each written independently**
  - Some which may do certain things better than others
  - All of which take time from their respective labs
  - None of which are collaborating with each other to present the ideal solution
  - May not even be publicly available for other researchers to use

- **Due to their status as a researcher in primarily another field**
  - Cannot provide ongoing support and are unwilling to bugfix for others
  - Abandon development once there is no longer incentive to (aka when the analysis is complete)

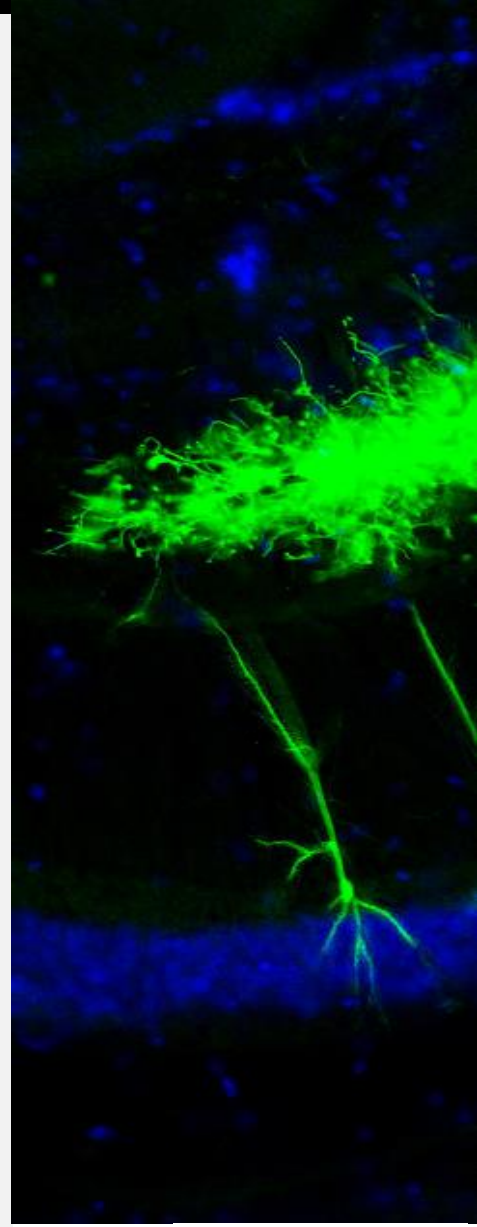


THEREFORE

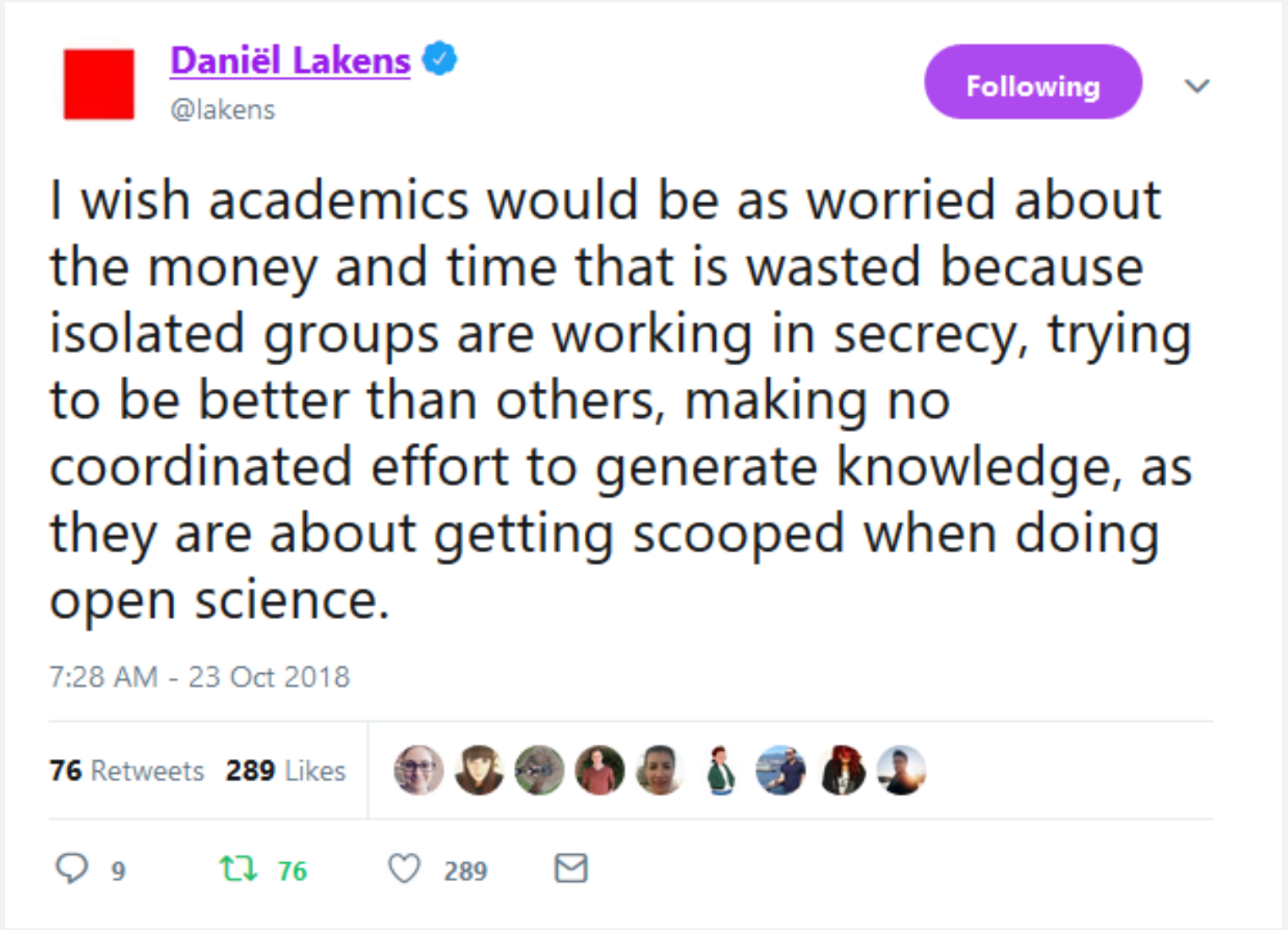
- You get many people writing similar code that does the same thing.

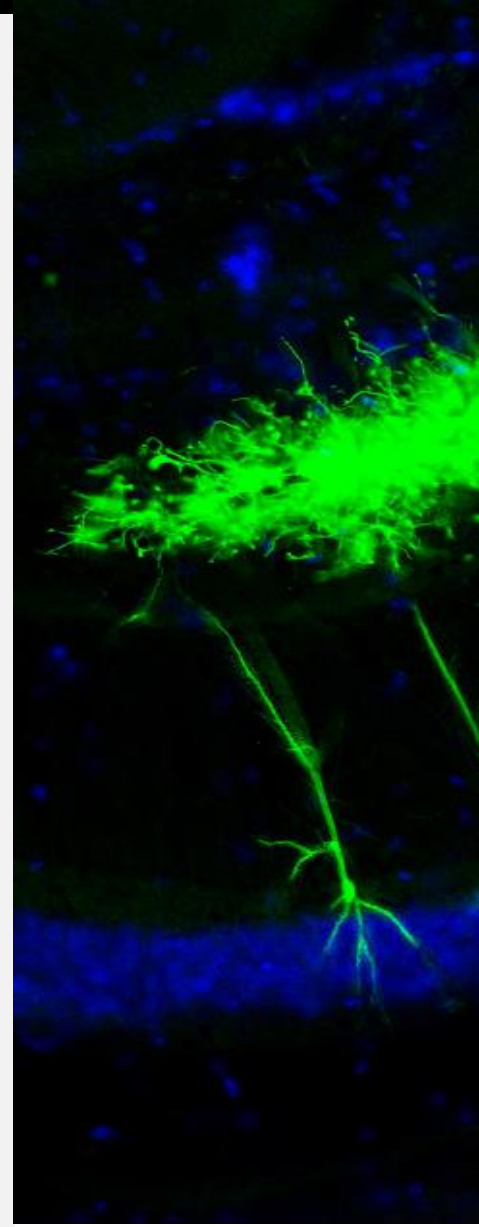
- Core issues:
  - High amounts of redundancy
  - Variability in quality
  - Not reproducible
  - May actually be buggy, especially across different platforms
  - Has no dedicated development team or community support for maintenance
  - **Not officially released**

- If it has none of the issues listed above, it is packaged and sold with a nice GUI
  - This is because most researchers not in computer science fields have no coding background and are thus code-phobic
  - Get those sweet, sweet citations
  - Can charge for licensing



THE ANTISOCIAL STATE OF ACADEMIA: A TWEET





@lisancao

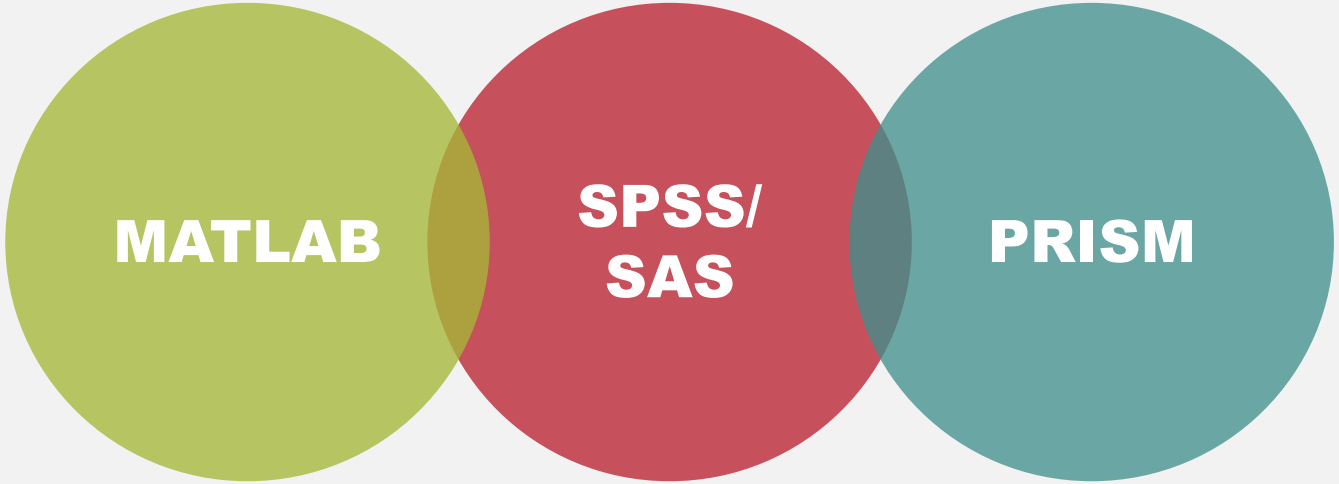
CENTRALIZATION & STRATIFICATION

To first de-centralize, we must first centralize

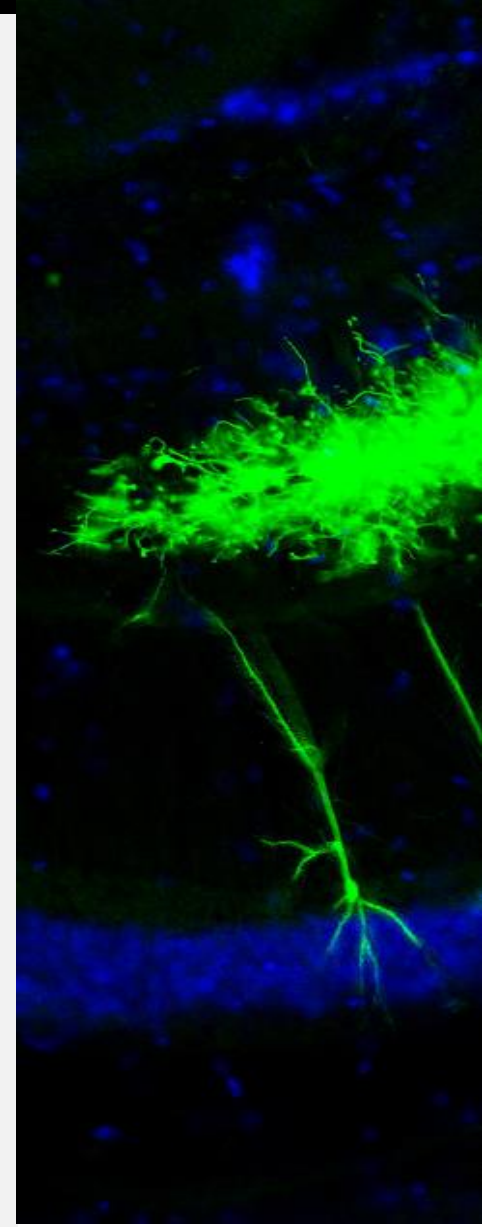
@lisancao

PROPRIETARY SOFTWARE IS HARMFUL TO SCIENTIFIC PROGRESS AND REPRODUCIBILITY

Domains of software: Getting stuck in your own bubble



"Well, if we already know how to use x software then we should just try to use it for everything, the university has already paid the licensing fee and we spent all this time learning it"



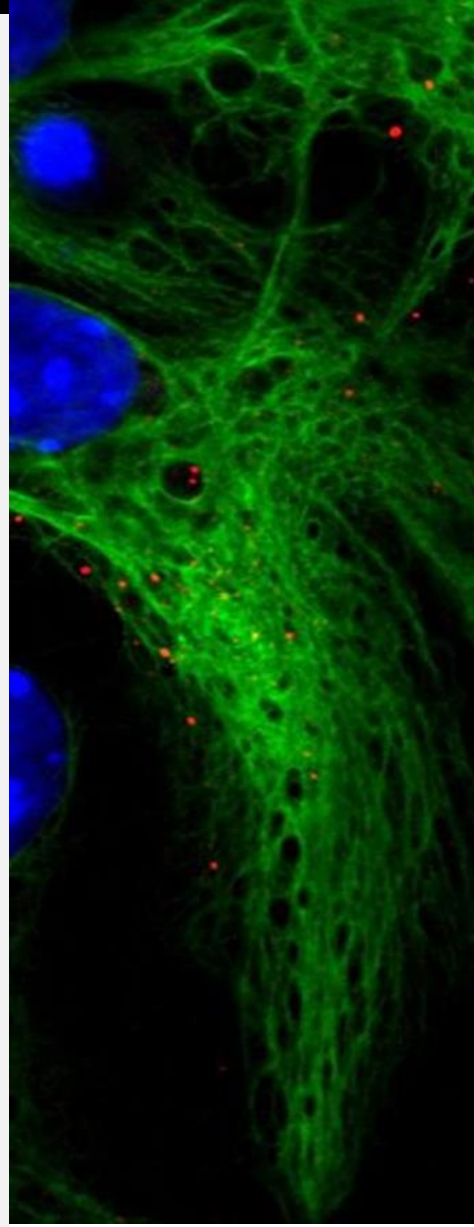
@lisancao

HOW CAN WE SOLVE THIS?

- **Centralization : continue building onto and prioritize educating researchers in existing programming languages that are free to use**
  - Build a community on a platform that fosters communication
  - Host on popular sites such as GitHub, GitLab and be as accessible as possible
  - **Improve the source code as much as possible and allow others to do the same, contribute to other projects and their source code**

- **Stratification: allow others to use your code, and actively use other people's code**
  - Understand what is useful and what isn't useful for what you are trying to do
  - Actively try to build comprehensive projects instead of patch-work scripts
  - Share your specialized version of the code with others

But most importantly: shift our attitudes and begin seeing tech proficiency as part of conducting cutting edge science



EXAMPLES OF STRATIFICATION AROUND A CENTRAL BASE CODE OR ALGORITHM

- Google's Research Division (Artificial Intelligence) <- NOT OPEN SOURCE!
  - Literally has scrapped all research projects that AI cannot be applied to
  - Focusses solely on building the best AI possible and applying it in as many ways possible

- The Linux Kernel (Stratified Distributions)

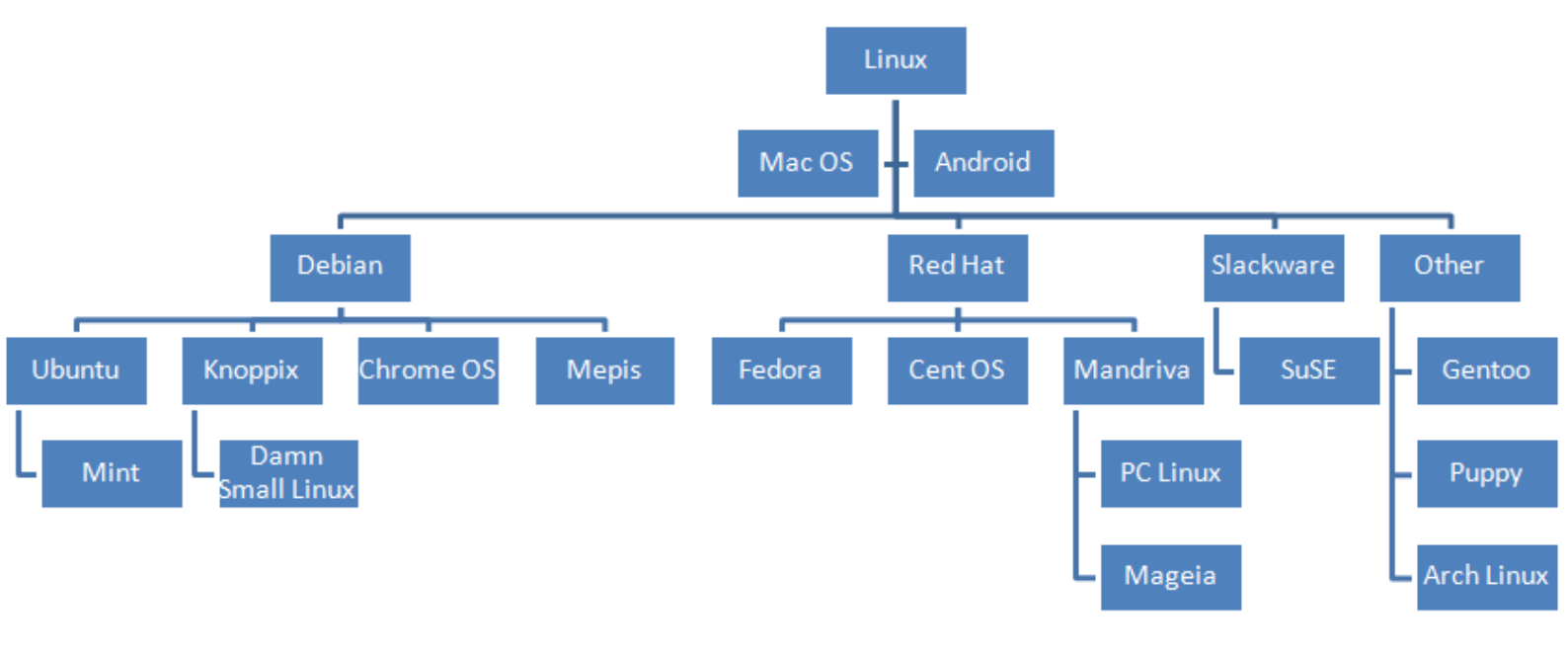
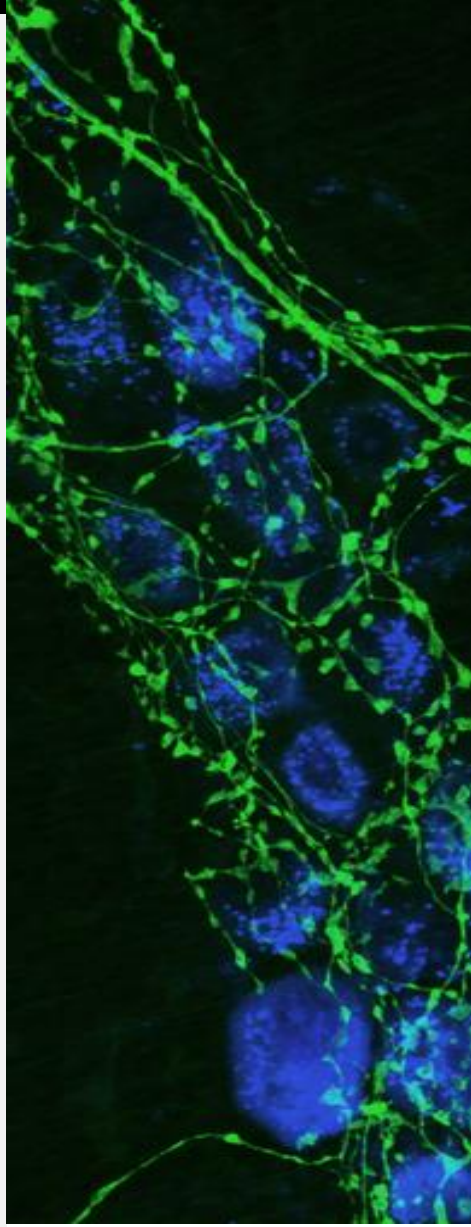


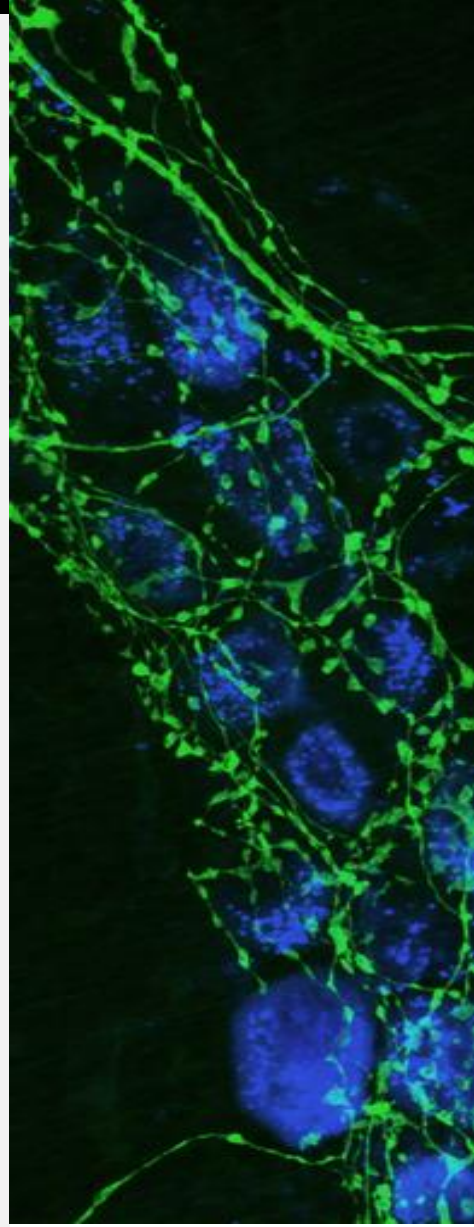
Image: Tellmeinsimpleterms.wordpress.com



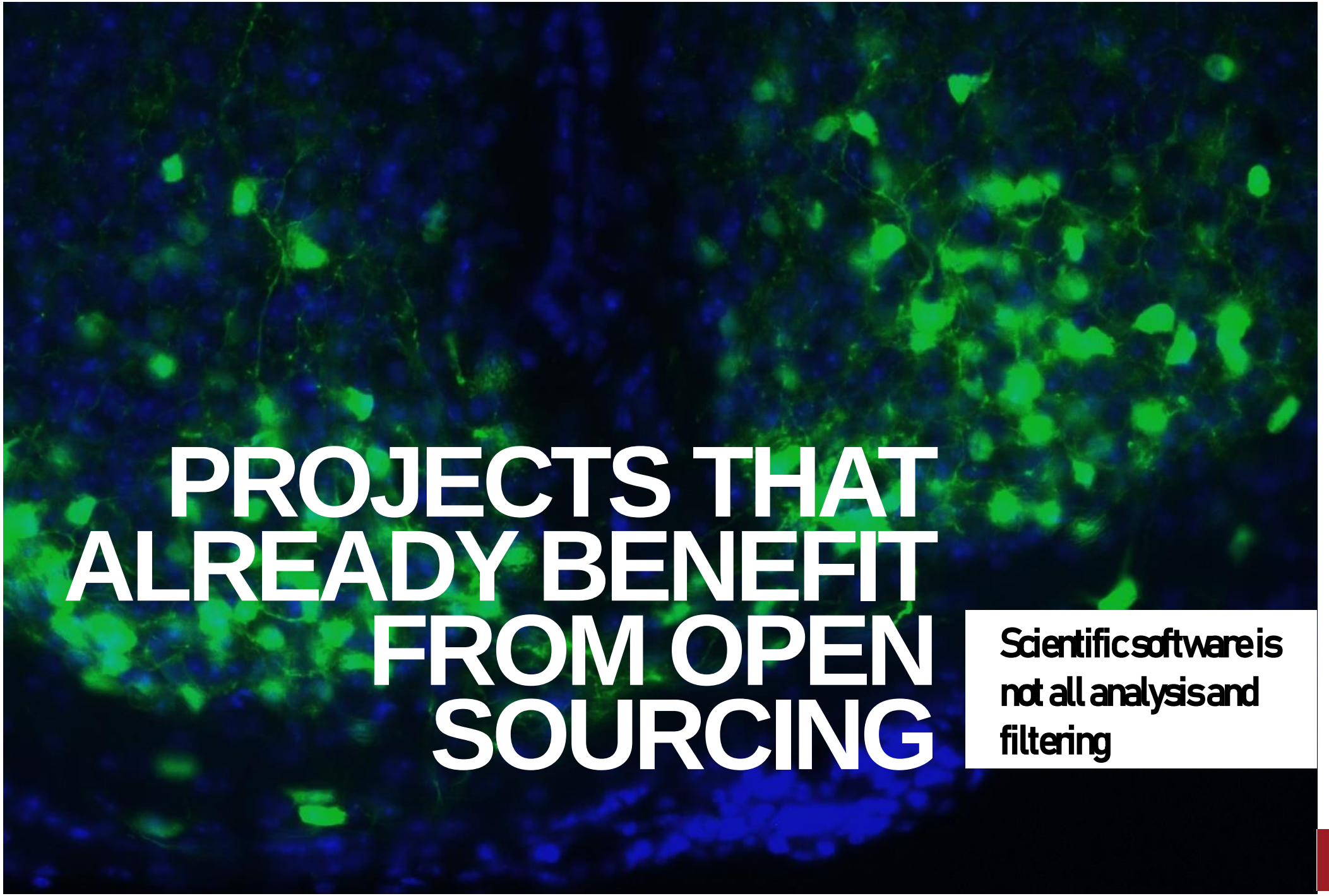
@lisancao

AN OPEN SOURCE FIELD: BIOINFORMATICS

- Basically genetics and computer science, this is one of the larger fields outside of traditional data science to adopt the open source model– many use a both Python and R as well as C/C++, Perl, Unix
  - R: Bioconductor base + supplemental packages, huge online community
  - Python: Biopython
- Often times more than one language is needed to develop your analysis pipeline, however both languages have packages that can work with each other (for example Python can be used in R, etc).
- Both languages have an extremely active community that embrace open source
- Each language has its strengths and weaknesses however there is nothing holding you back from simply using both



@lisancao



PROJECTS THAT ALREADY BENEFIT FROM OPEN SOURCING

Scientific software is not all analysis and filtering

@lisancao

OPENWORM (2014)

A project that started off as a Kickstarter fundraiser, is an independent science project that sought to perfectly emulate the C. Elegans worm on personal computers.

-The first complete simulation of an entire organism

-Raised $121,076 with 799 backers

-Is a significant building block to eventually emulating a rodent model, which is much more complex. This amount of work would no doubt require the expertise and hard work of hundreds of people– ideal for open sourcing.

From OpenWorm we have the following projects:

-OpenWorm Browser, DevoWorm, NeuroML Connectome, Geppetto, Sibernetic (Worm body simulator), ChannelWorm2, c302, WormSim, & more

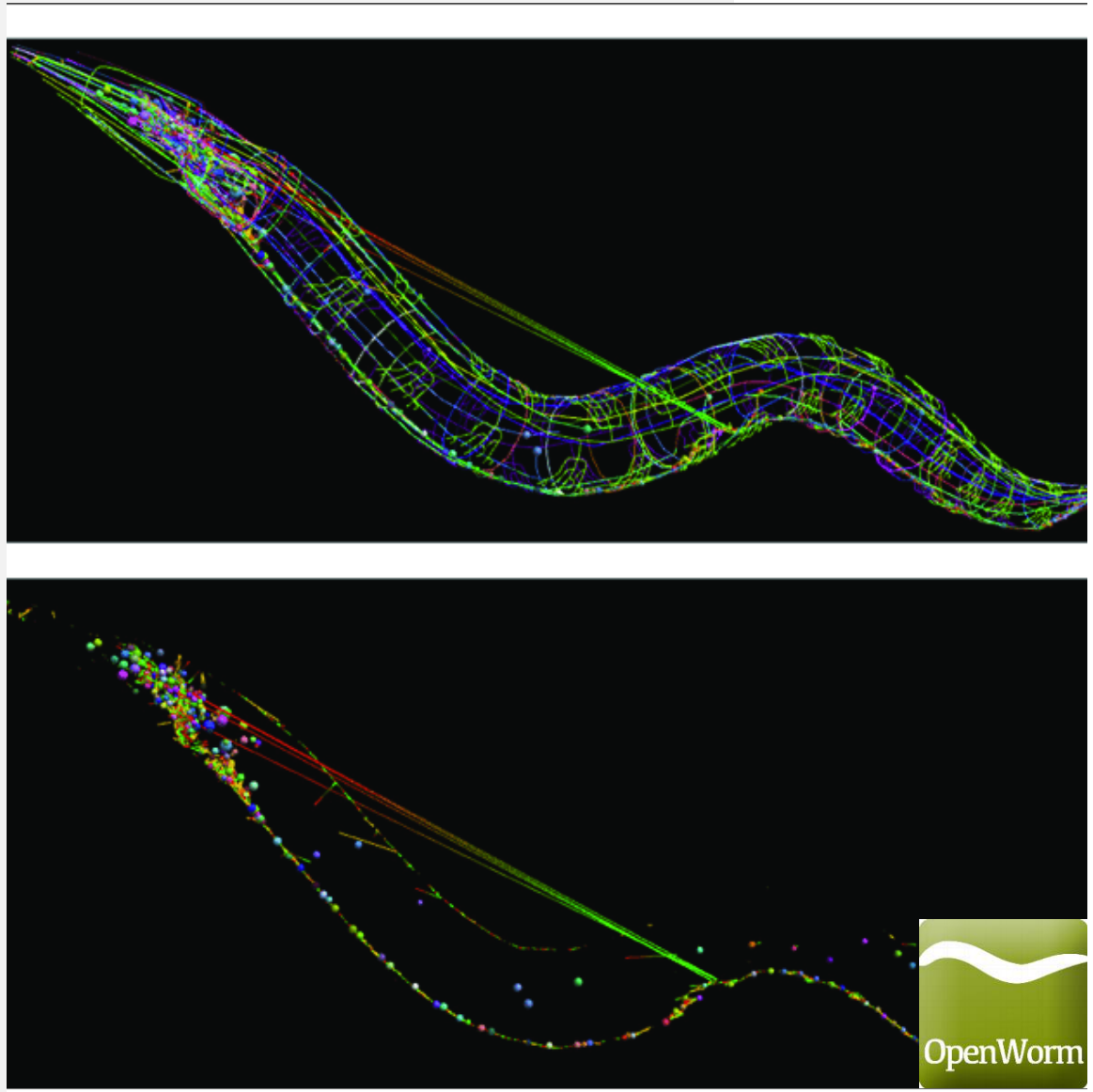


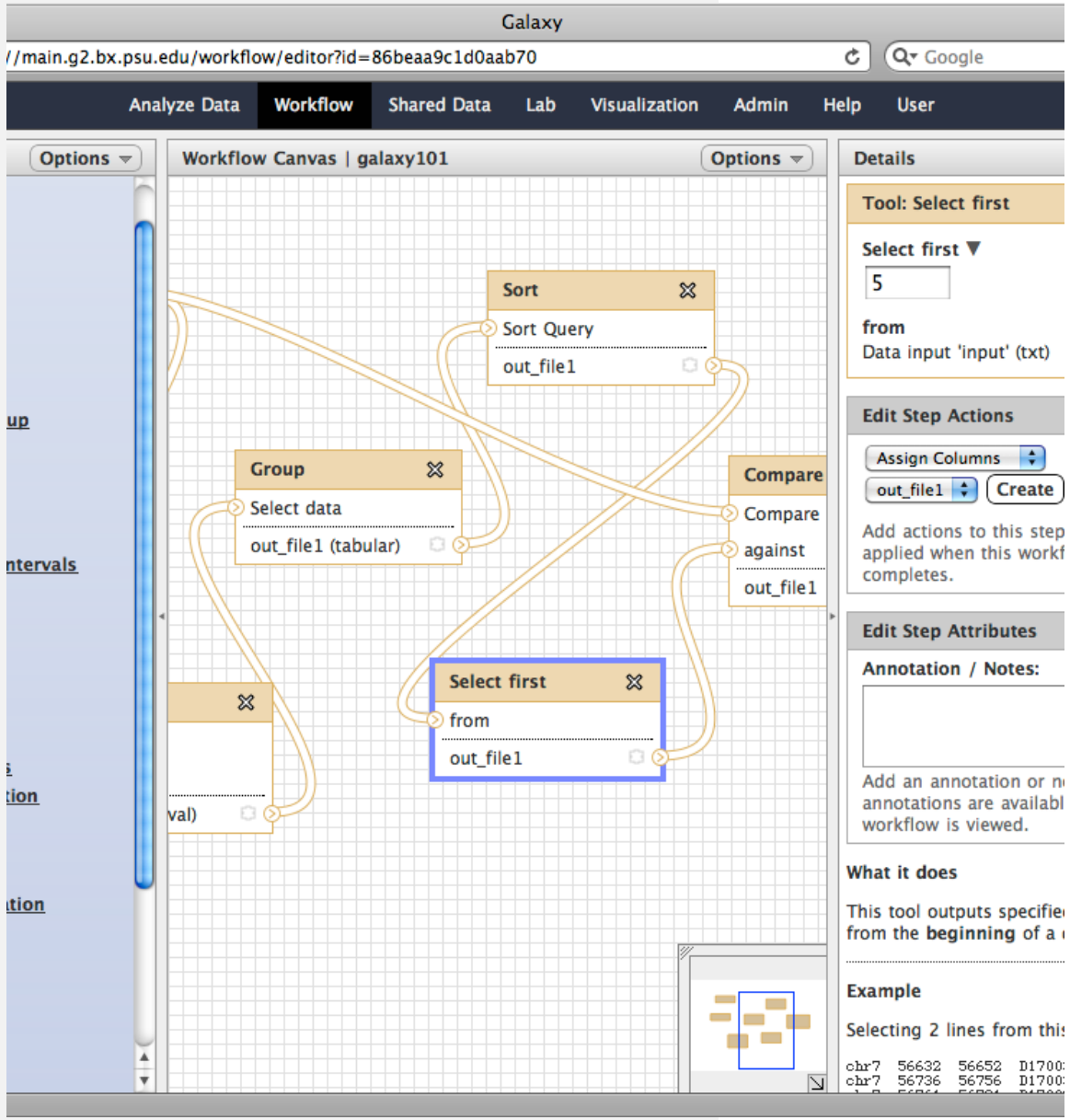
Image: Vella, Mike et al.(2014). libNeuroML and PyLEMS: Using Python to combine procedural and declarative modeling approaches in computational neuroscience. Frontiers in neuroinformatics. 8. 38. 10.3389/fninf.2014.00038.

https://github.com/openworm

GALAXY PROJECT

"Galaxy is a scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming or systems administration experience."

- GUI friendly but still open source workflow intended to be used by researchers who are using bioinformatics in their work but are not trained as bioinformaticians
- By actively developing and fostering a great community, the users don't have to worry about the under-the-hood of the application
- Specialized tools are developed for most cell types and different analyses for various data types
- Makes a cutting edge field rather accessible



https://usegalaxy.org

Galaxy PROJECT

OPEN A.I. PROJECT

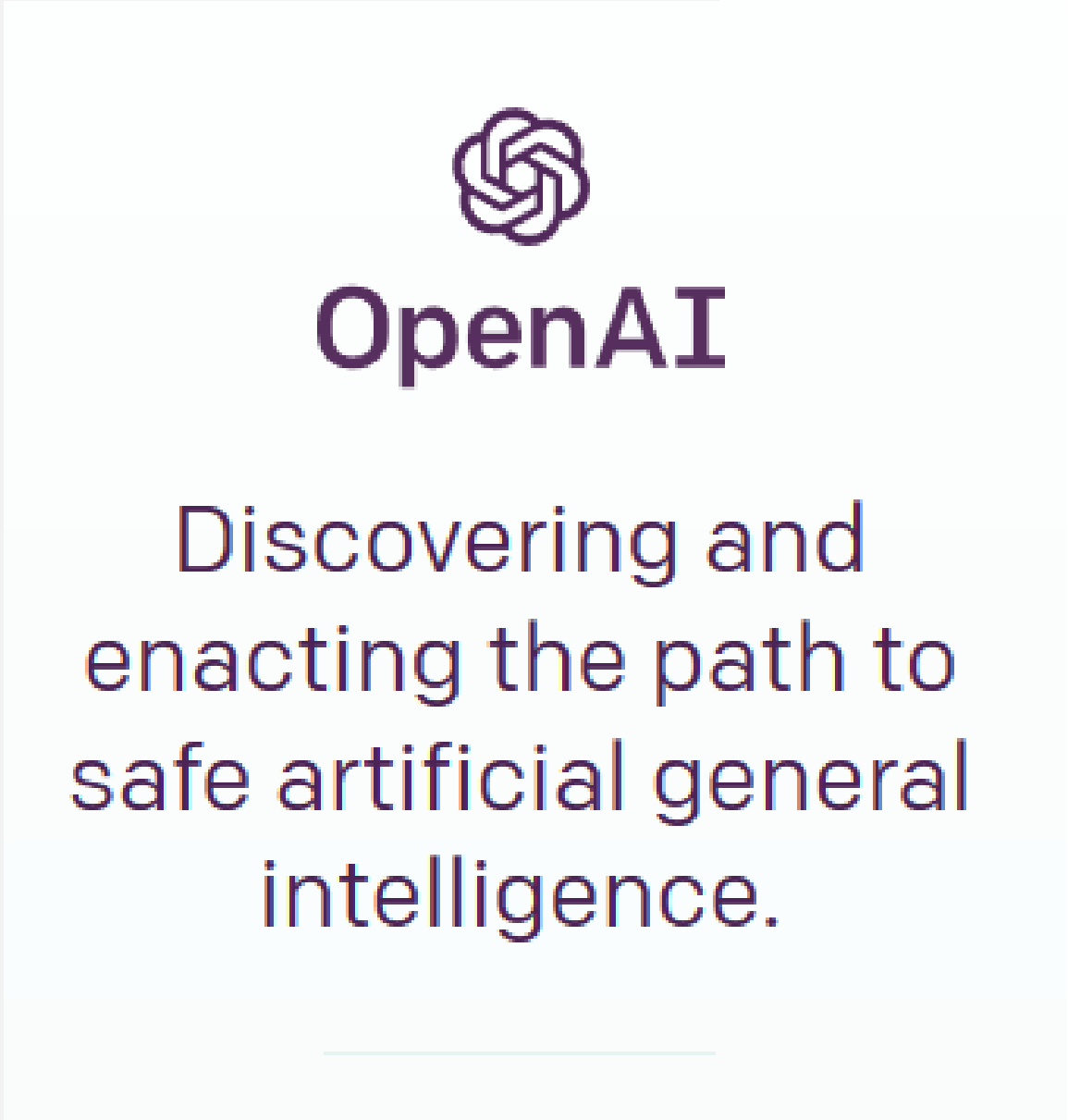
"OpenAI is a non-profit AI research company, discovering and enacting the path to safe artificial general intelligence."

-Many researchers are interested in incorporating AI or Machine Learning into their work, but lack the knowledge needed to make their own algorithms

-OpenAI already conducts research and publishes regularly, and provides a means towards the safe development of AGI

-Currently staff 60 researchers and engineers full-time

-Develop open source platforms and tools for researchers to use and implement algorithms into their work

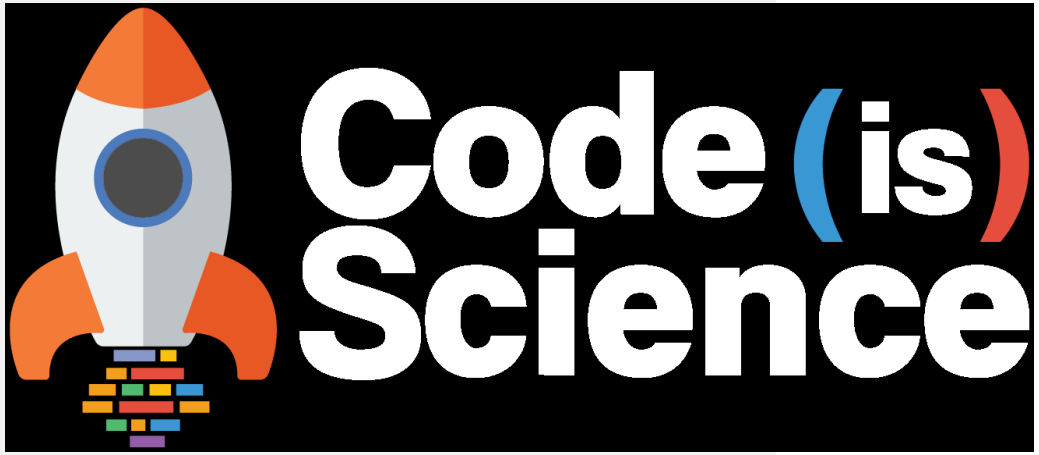


https://github.com/openai

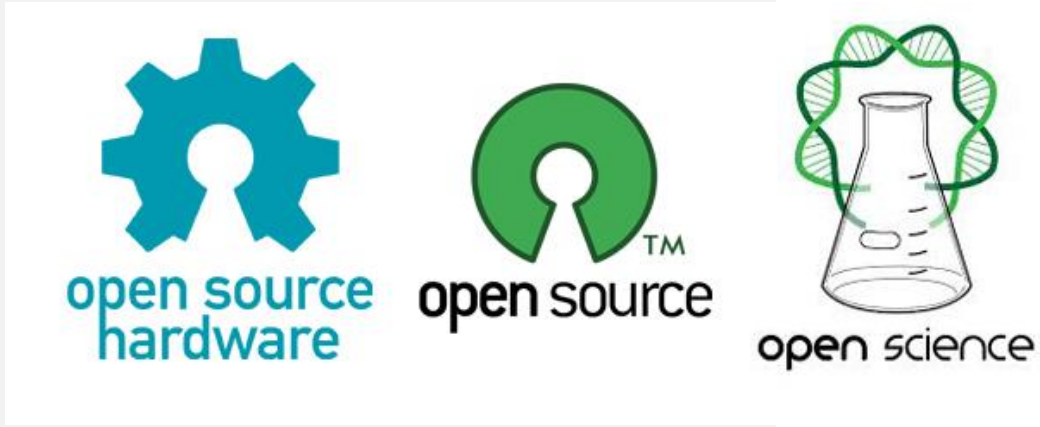
OTHERS


OPEN BIOMEDICAL
INITIATIVE


OpenMD


Code (is) Science


open
source
imaging


open source
hardware
open source
TM
open science


OPEN BEHAVIOR


SciPy.org


OpenLabTools

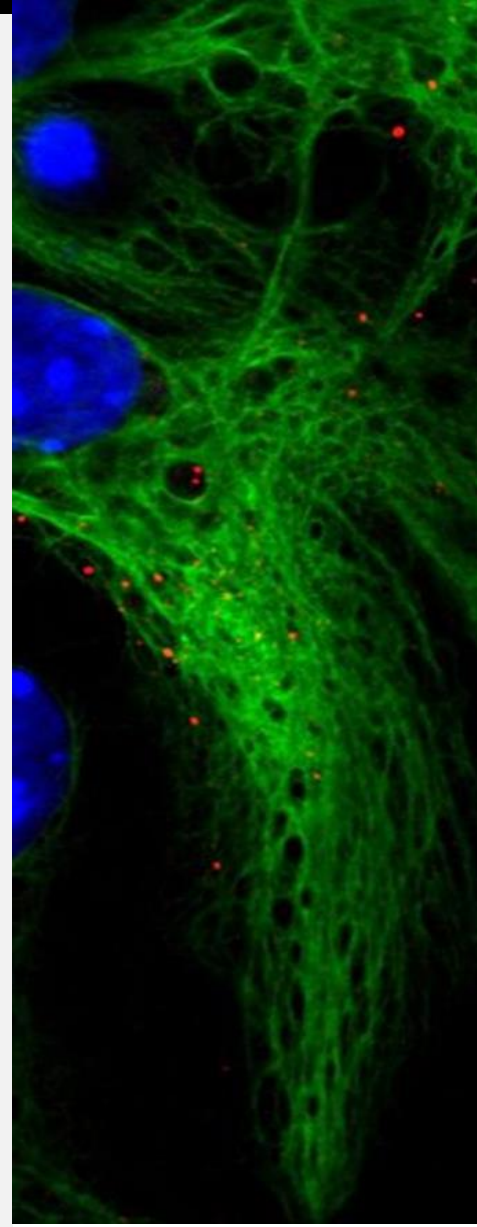

software carpentry


OPEN SOURCE
DRUG DISCOVERY


NIH
3D PRINT
EXCHANGE

IN SUMMARY: WHY SHOULD WE OPEN SOURCE SCIENTIFIC SOFTWARE?

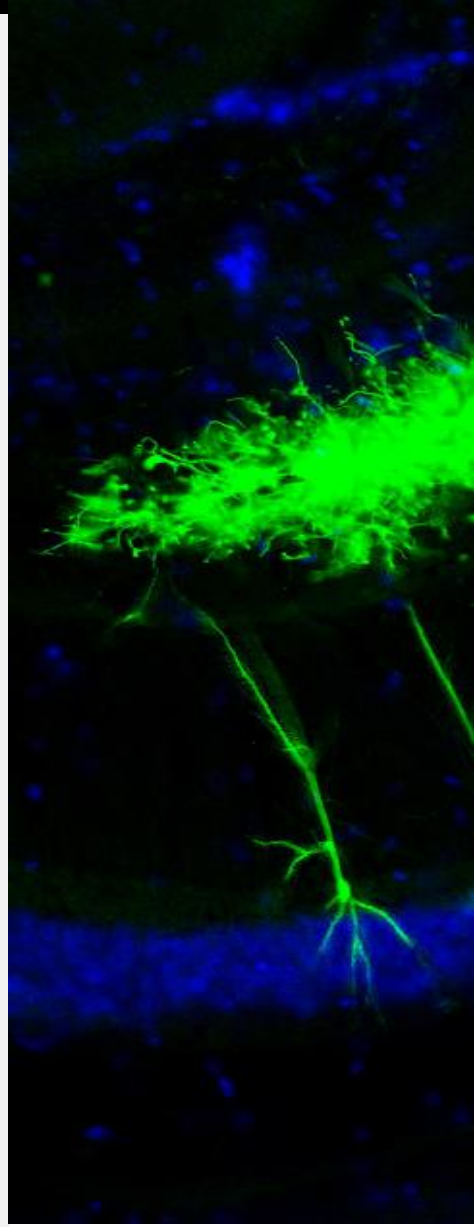
1. Reproducibility
   - Without the original code, nobody will ever be able to truly reproduce the results
2. Transparency
   - If the code used for analysing the data was poorly written, we need to know
3. Contributability
   - If the code is bad, let others help you improve it!
4. Reduced Costs
   - Let's cut down licensing costs on proprietary software
5. Reduced Redundancy
   - Why waste time writing code when it already exists and is actively maintained?
6. Better Tools to Streamline Innovation
   - This way we are able to take full advantage of new technology as it rolls out and are not reliant on a core development team to implement them into their own software



HONEST THOUGHTS

What seems to be the main issue here is not that there aren't enough benefits for open sourcing, but instead that our holdbacks are too great

- Our institutions needs to incentivize and reward contributing to open source, actively maintaining projects, and bug fixing
- Tenure-track prospects must go beyond just impact scores and citation counts
- Community engagement needs to be recognized as contributing to science, and in a big way
- Researchers need to connect with each other online and troubleshoot openly
- There needs to be a way to directly streamline programmers and computer scientists into projects from fields that don't have strong coding backgrounds
- We need to end the hold current propriety software has on universities



A BIG THANK YOU TO:

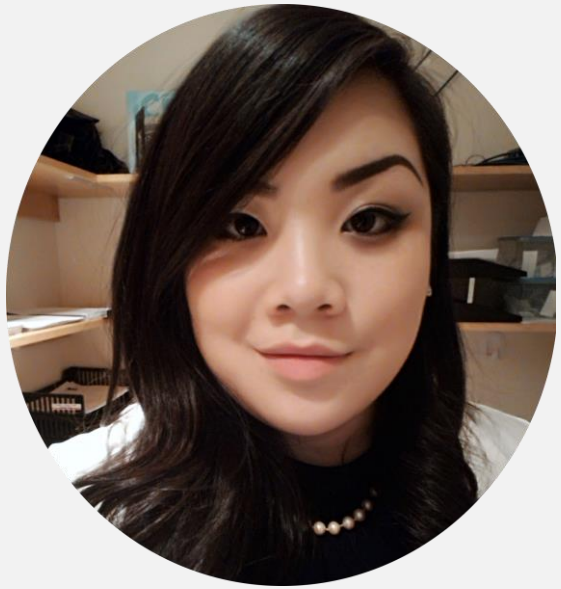
pyladies

ABOUT ME

Research Assistant at the Circadian Rhythms and Sleep Neuroscience Lab & Behavioural Neuroendocrinology Lab, SFU Psychology Department.

SFU Scientific Programming Group

RLadies & PyLadies Member



FIND ME ON SOCIAL MEDIA:

- *Twitter: @lisancao*
- *lisanatashacao@gmail.com*
- *Github: github.com/lisancao*