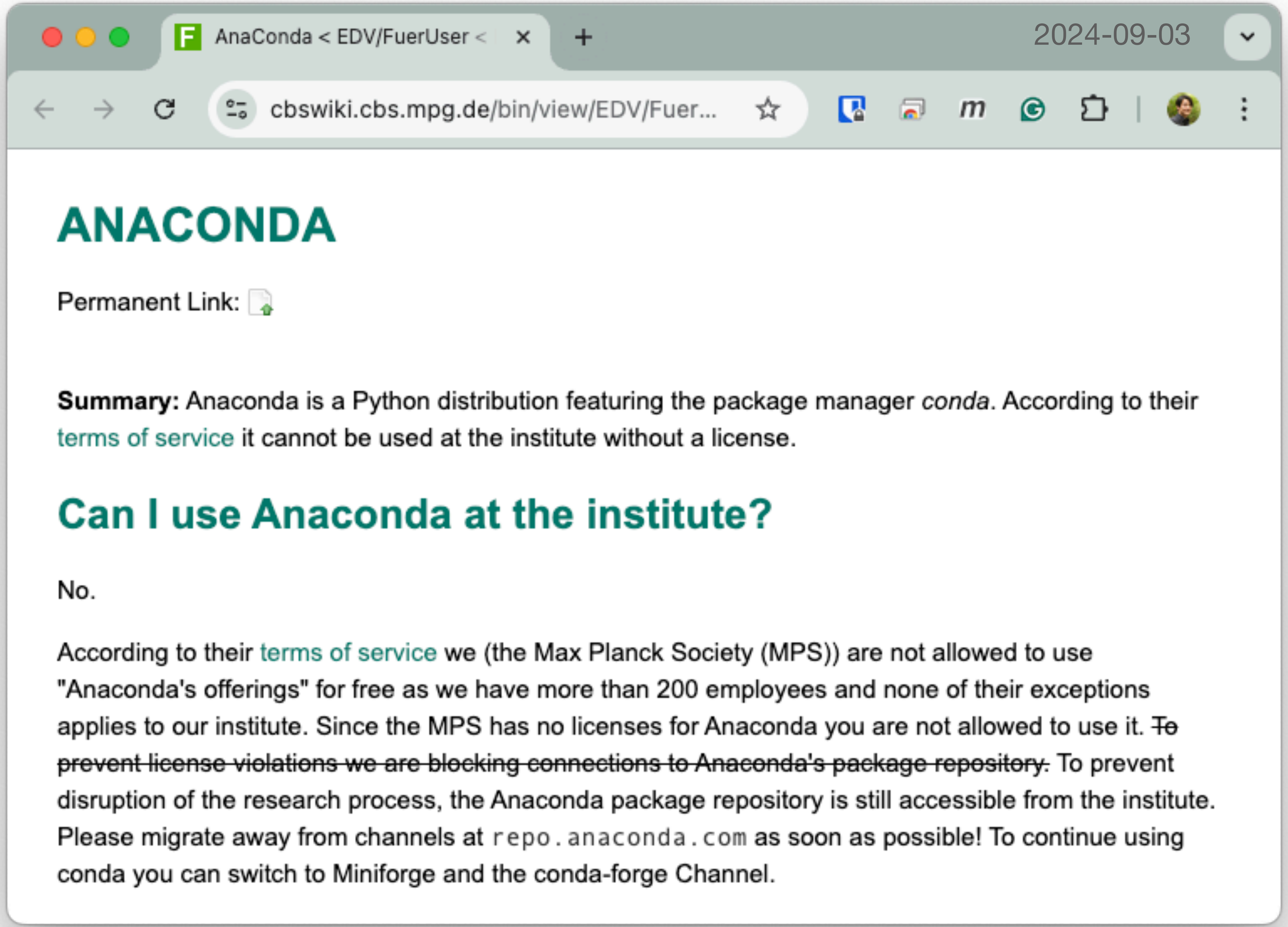
Open Questions for Computing Language for Open Science

NCML lab meeting

2024-10-02 Seung-Goo KIM

Motivation

Anaconda drama



AnaConda < EDV/FuerUser <
2024-09-03
cbswiki.cbs.mpg.de/bin/view/EDV/Fuer...

ANACONDA

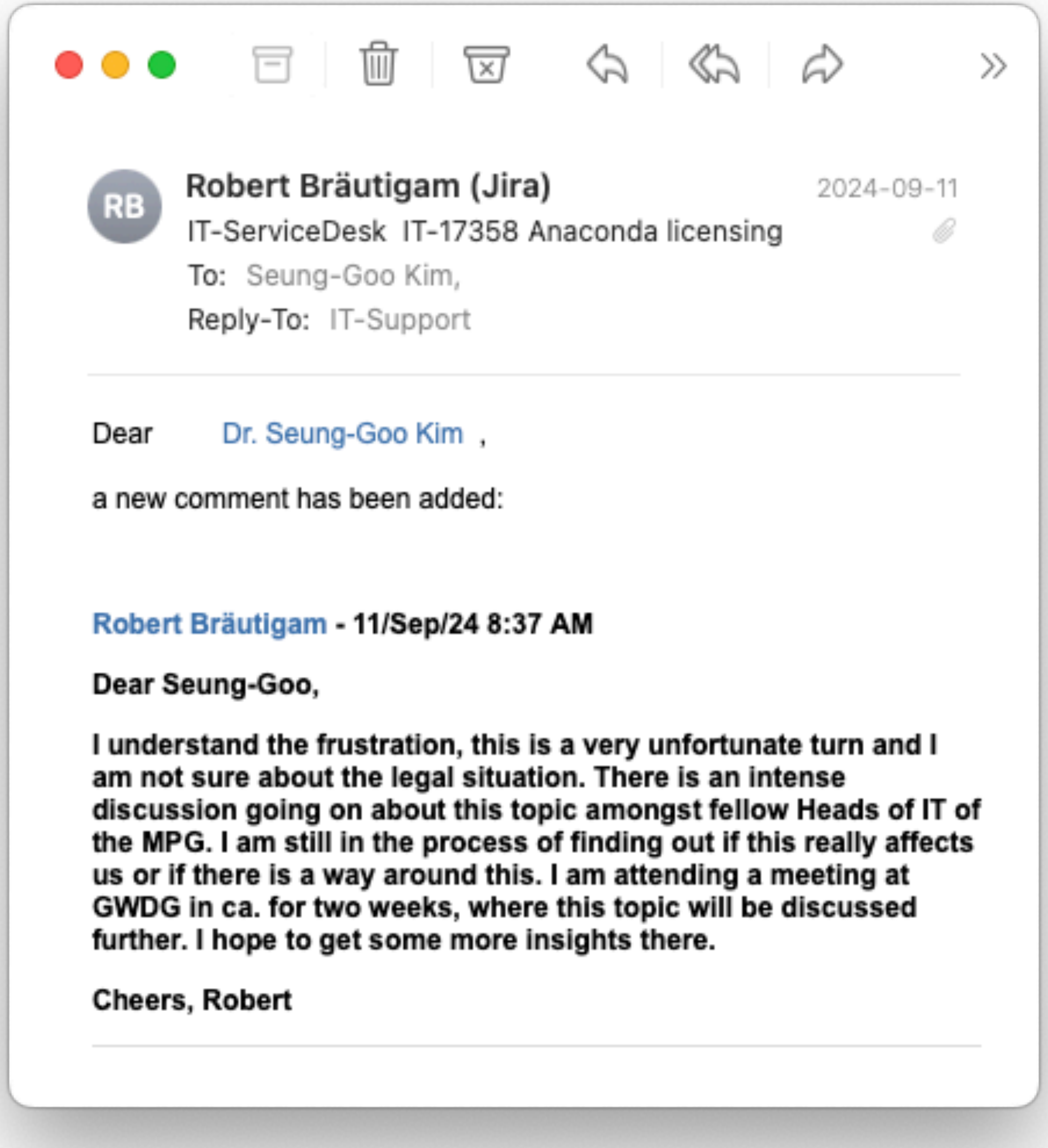
Permanent Link:

Summary: Anaconda is a Python distribution featuring the package manager *conda*. According to their terms of service it cannot be used at the institute without a license.

Can I use Anaconda at the institute?

No.

According to their terms of service we (the Max Planck Society (MPS)) are not allowed to use "Anaconda's offerings" for free as we have more than 200 employees and none of their exceptions applies to our institute. Since the MPS has no licenses for Anaconda you are not allowed to use it. ~~To prevent license violations we are blocking connections to Anaconda's package repository.~~ To prevent disruption of the research process, the Anaconda package repository is still accessible from the institute. Please migrate away from channels at `repo.anaconda.com` as soon as possible! To continue using conda you can switch to Miniforge and the conda-forge Channel.



Robert Bräutigam (Jira) 2024-09-11
IT-ServiceDesk IT-17358 Anaconda licensing
To: Seung-Goo Kim,
Reply-To: IT-Support

Dear Dr. Seung-Goo Kim ,

a new comment has been added:

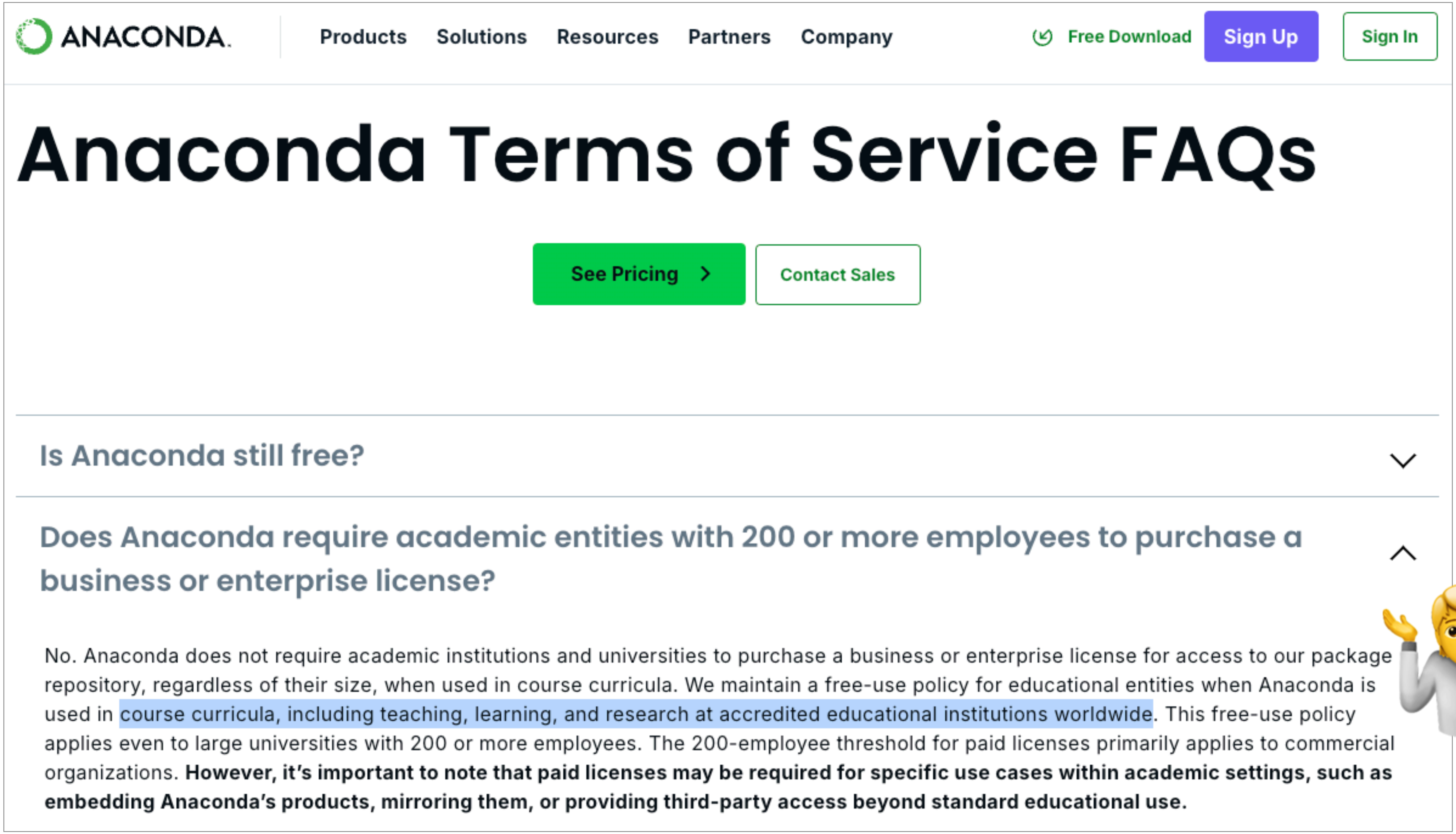
Robert Bräutigam - 11/Sep/24 8:37 AM

Dear Seung-Goo,

I understand the frustration, this is a very unfortunate turn and I am not sure about the legal situation. There is an intense discussion going on about this topic amongst fellow Heads of IT of the MPG. I am still in the process of finding out if this really affects us or if there is a way around this. I am attending a meeting at GWDG in ca. for two weeks, where this topic will be discussed further. I hope to get some more insights there.

Cheers, Robert

https://www.anaconda.com/pricing/terms-of-service-faqs [2024-10-01]



Anaconda Terms of Service FAQs

Is Anaconda still free?

Does Anaconda require academic entities with 200 or more employees to purchase a business or enterprise license?

No. Anaconda does not require academic institutions and universities to purchase a business or enterprise license for access to our package repository, regardless of their size, when used in course curricula. We maintain a free-use policy for educational entities when Anaconda is used in course curricula, including teaching, learning, and research at accredited educational institutions worldwide. This free-use policy applies even to large universities with 200 or more employees. The 200-employee threshold for paid licenses primarily applies to commercial organizations. **However, it's important to note that paid licenses may be required for specific use cases within academic settings, such as embedding Anaconda's products, mirroring them, or providing third-party access beyond standard educational use.**

What computing language should we use?

What computing language should we use for open science?

Why do I associate Python with Open Science?

- Until 2000s?: We used proprietary software all the time (MATLAB, BESA, BrainVoyager, SPSS, ..., Microsoft Office, Windows, Mac{HW+SW})
- 2012: Special Section on "Replicability in Psychological Science" in Perspectives on Psychological Science, https://journals.sagepub.com/doi/10.1177/1745691612465253
- 2015: Open Science Collaboration, "Estimating the reproducibility of psychological science", Science, https://doi.org/10.1126/science.aac4716
- For **transparency**, the whole analysis needs to be replicable on others' system only using open-access software (i.e., docking everything won't violate any copyright laws) => "MATLAB is proprietary🤑. Thus, MathWorks is the enemy of OpenScience👹!"
- Now if we need to pay for Anaconda (or maybe Python🐍 itself), it would make switching to Python meaningless for Open Science!

Skepticism triggered! 🤔🧐🤨

A punch in the face wakening up from the naiveness

- Is every open-access project👹 just a long-term investment for the market dominance?
- Is every public talk😈 an advertisement for a book, a promotion for a product, or a propaganda for a vote?
- "*If you're not paying for the product, then you are the product.*" (Tristan Harris🧌) (or is he just selling his books?)
- (Am I scamming you right now?😨🙀)
- ...but let's get real and let's think about the problem a bit more

Why did I start using those languages?

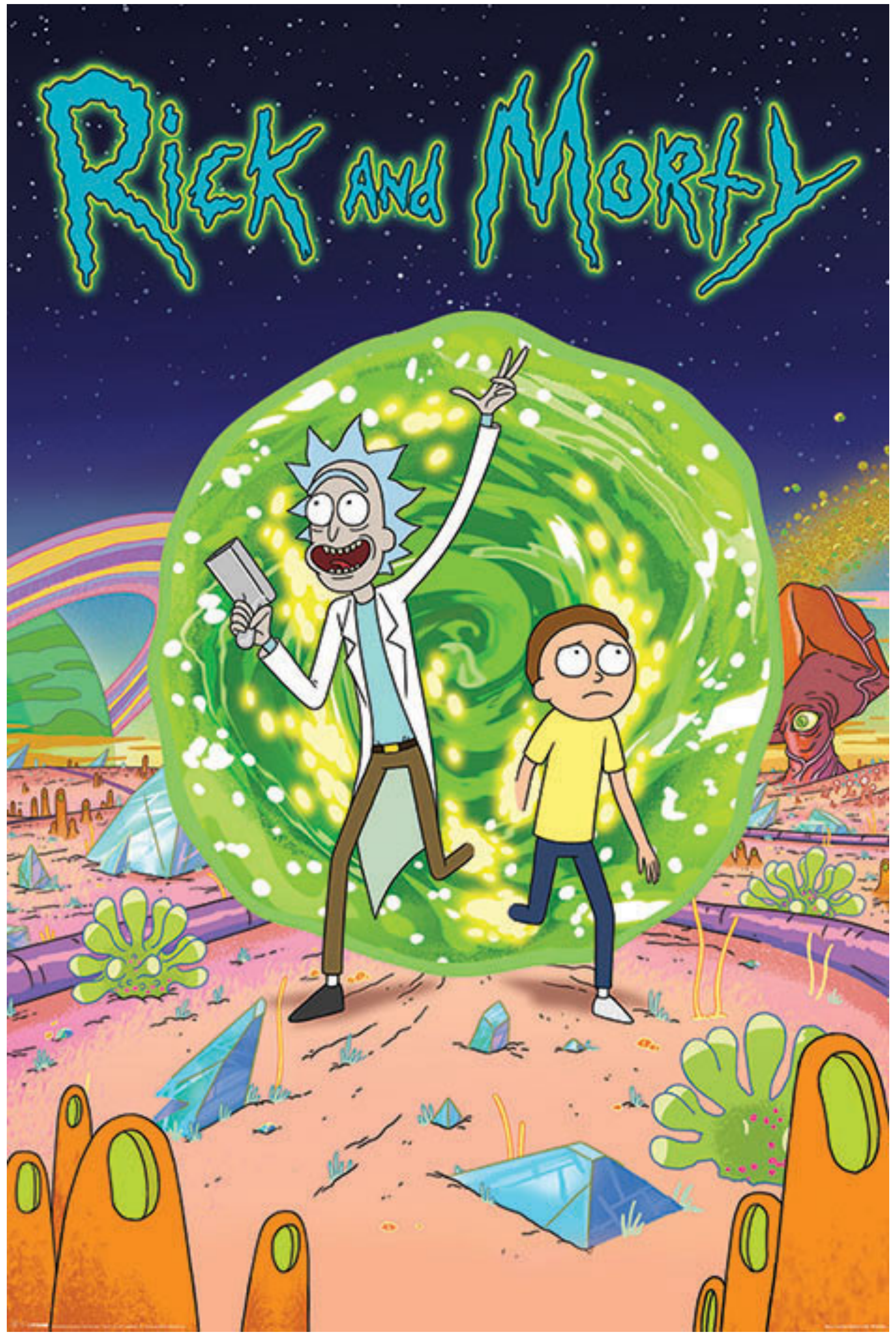
But first...

- **MATLAB**: because of SPM, SurfStat, EEGLAB, FieldTrip, and other popular toolboxes
- **Python**: because of MNE-python, TensorFlow, and other popular packages
- **R**: because of linear mixed-effects models (lme4)
- **Bash**: because of FSL, ANTs, FreeSurfer
- I didn't choose those language. I only chose packages.
- The developers of the popular packages/libraries chose their languages!
- So, how much meaningful is it for me to consider which language is good or bad?🤔

Are we end-users or developers?

How much helpful to learn languages?

- Some experimental physicists (like CERN or NASA) build their own tools [e.g., large hadron colliders, space telescopes, ...].
- We mainly develop theories, designs, and analyses, but also sometimes "design" new devices/environments [e.g., MR-piano, ArtLab].
- [my take]: Normally we use software as a tool, but sometimes it helps if we know how to tweak a little (it is also a bit risky though).



https://www.europosters.de/poster/rick-morty-portal-v35685

Open questions

- Is proprietary software evil?
- Can the transparency be implemented only by open-access software?
- What is the best language?

Open questions

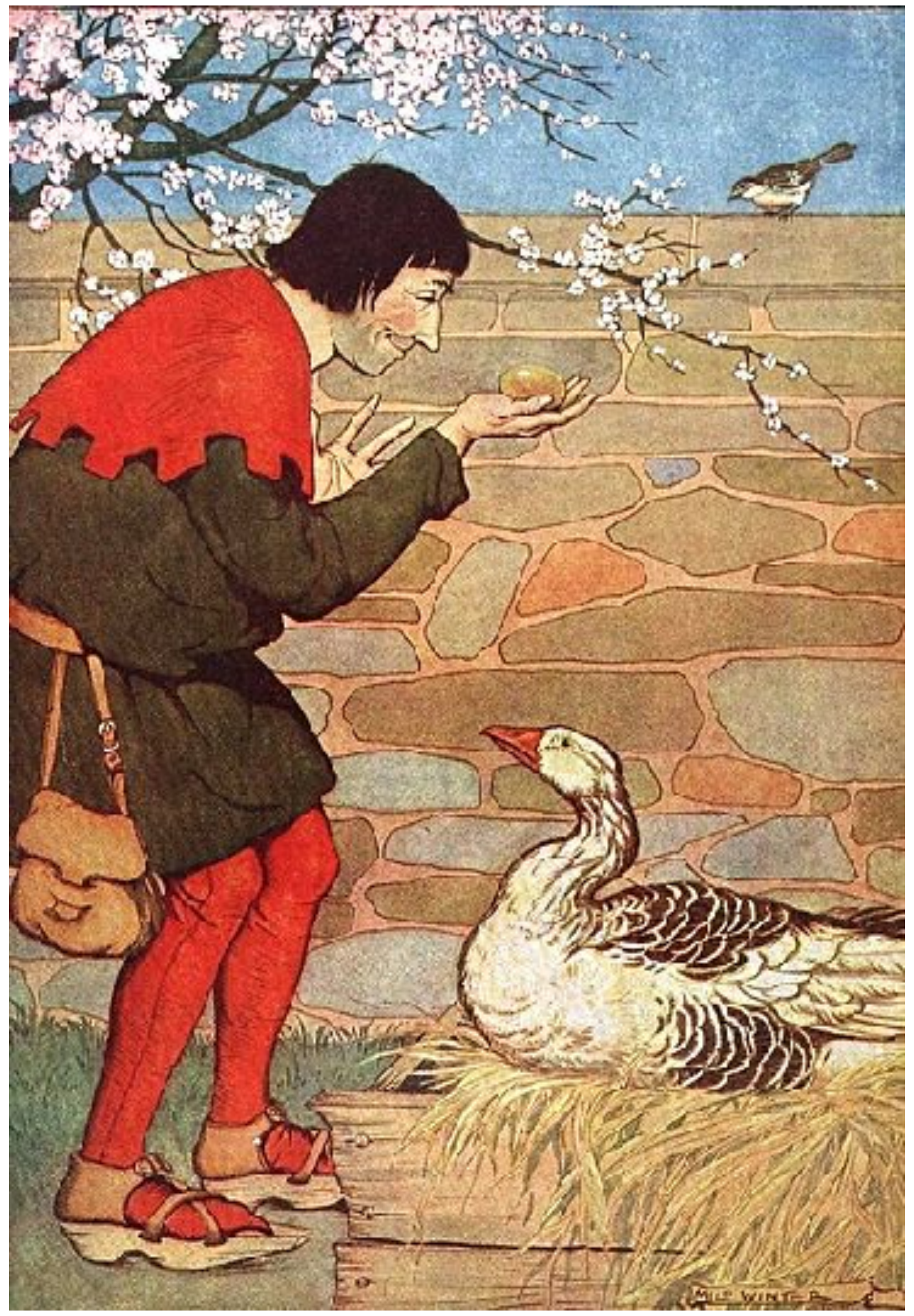
- Is proprietary software evil?
- Can the transparency be implemented only by open-access software?
- What is the best language?

Is capitalism inherently evil?

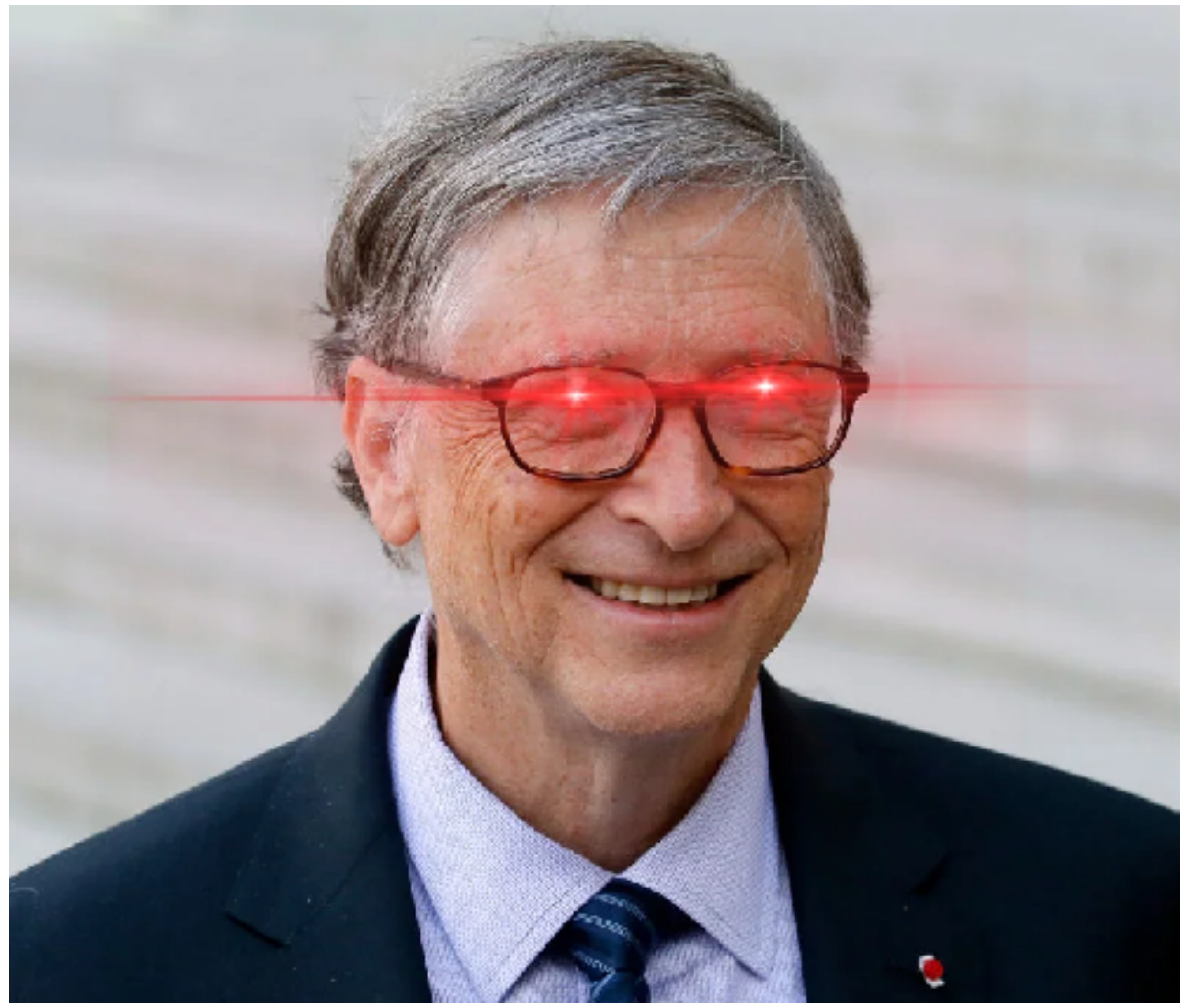
Why some people feel software should be ethically free?

Those who have plenty want more and so lose all they have.

St. IGNUcius, the Church of EMACS (2012)
Free Software Foundation (1985-)



The Goose that Laid the Golden Eggs
Aesop's fable (~600 BCE), wiki



Bill Gates with red eyes, r/linuxmemes

How do they make open access software?

Are they just heavenly creatures living off their own niceness?

- **Linus Torvalds** (the creator of Linux) gets 1.7M USD in compensation as a Fellow of Linux Foundation
- **Guido van Rossum** (the creator of Python) worked at own startups, and Google[05-12], Dropbox[13-19], Microsoft[20-], 1-5M USD/yr?
- **Ross Ihaka** & **Robert Gentleman** (the co-creators of R) worked as statistic professors at University of Auckland, NZ
- **Travis Oliphant** (the creator of NumPy, SciPy and a co-founder of NumFOCUS, Anaconda[CEO:12-17]) gets compensation from as a staff at NumFOCUS.
- **Bjarne Stroustrup** (the creator of C++) worked as a programmer at Bell Labs, as a professor at Texas A&M University, and Columbia University.

When you see a nice open-access project...

How long will it be maintained and last open-access?

- A non-profit foundation?
- A for-profit start-up?
- A fixed-term research funding?
- A tenured professor's hobby?

Is software public goods or common goods?

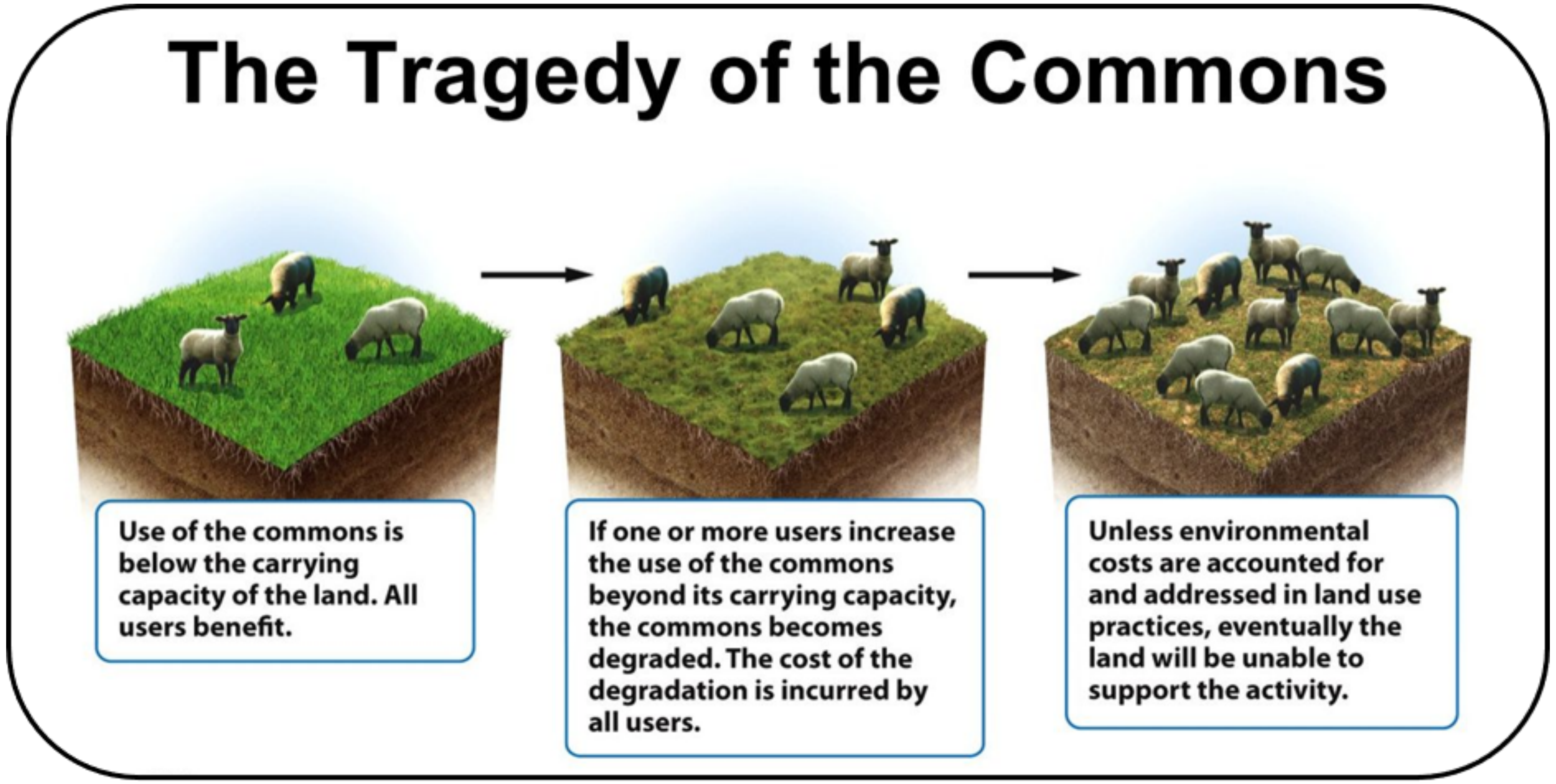
- Excludability: a possibility to exclude access of non-owners
- Rivalry: one's usage does not affect the availability for others
- Goods:

	Excludable	**Non-excludable**
Rivalrous	Private goods	Common goods (e.g., open data?)
Non-rivalrous	Club goods (e.g., proprietary s/w)	Public goods (e.g., open source s/w)

Common goods

In capitalistic economics, it's called "externality"

- Hardin G, 1968, The Tragedy of the Commons, *Science*.
    - Depletion of public resources is inevitable (private or public ownership is needed)
- Dietz et al., 2003, The Struggle to Govern the Commons, *Science*.
    - Can be avoided by communal efforts and other institutions (i.e., making them club goods)



The Tragedy of the Commons

Use of the commons is below the carrying capacity of the land. All users benefit.

If one or more users increase the use of the commons beyond its carrying capacity, the commons becomes degraded. The cost of the degradation is incurred by all users.

Unless environmental costs are accounted for and addressed in land use practices, eventually the land will be unable to support the activity.

https://www.sustainable-environment.org.uk/Earth/Commons.php

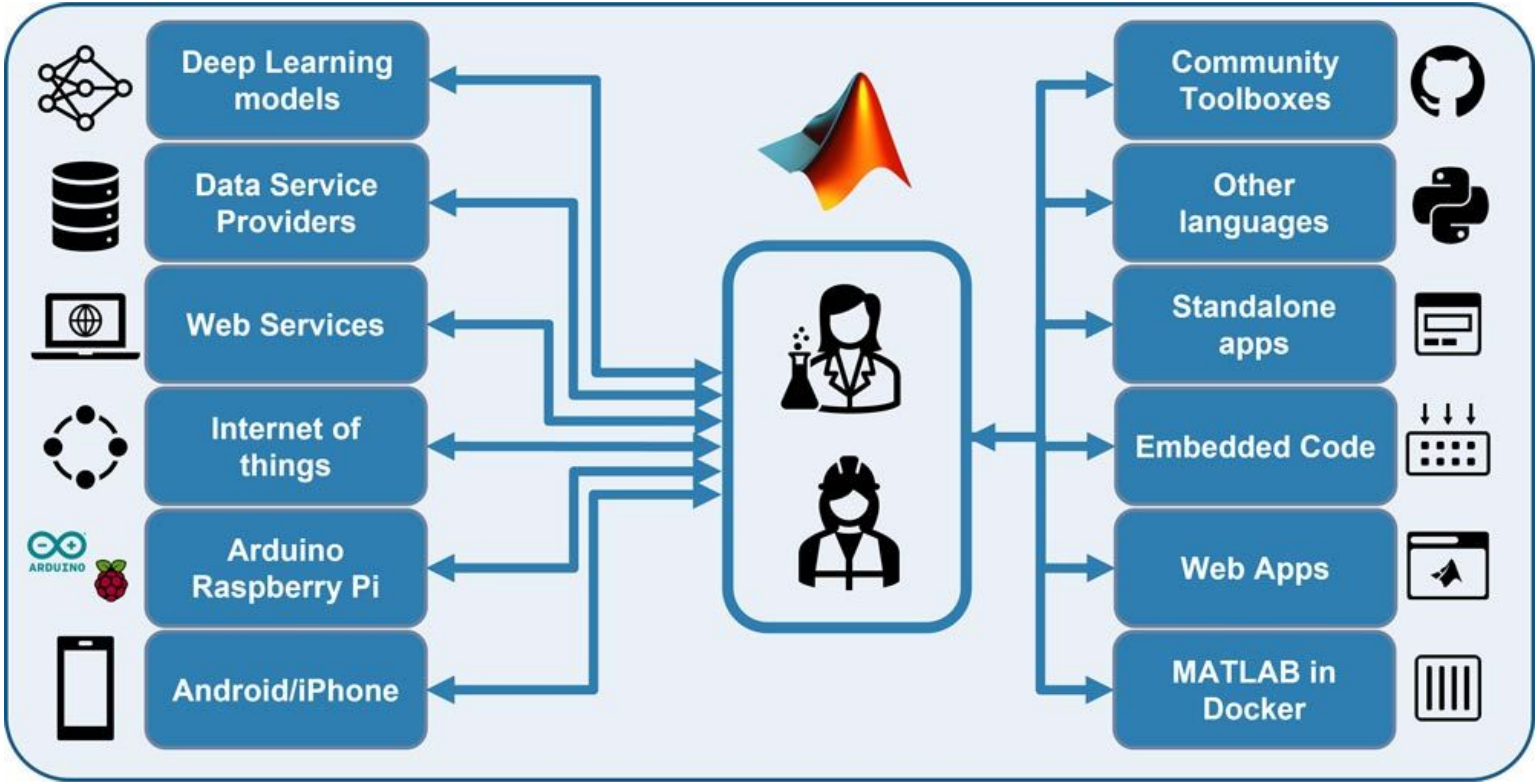
Your thoughts? 🧐💭

Open questions

- Is proprietary software evil? (should/can we be non-evil in the long run?)
- Can the transparency be implemented only by using open-access languages?
- What is the best language?

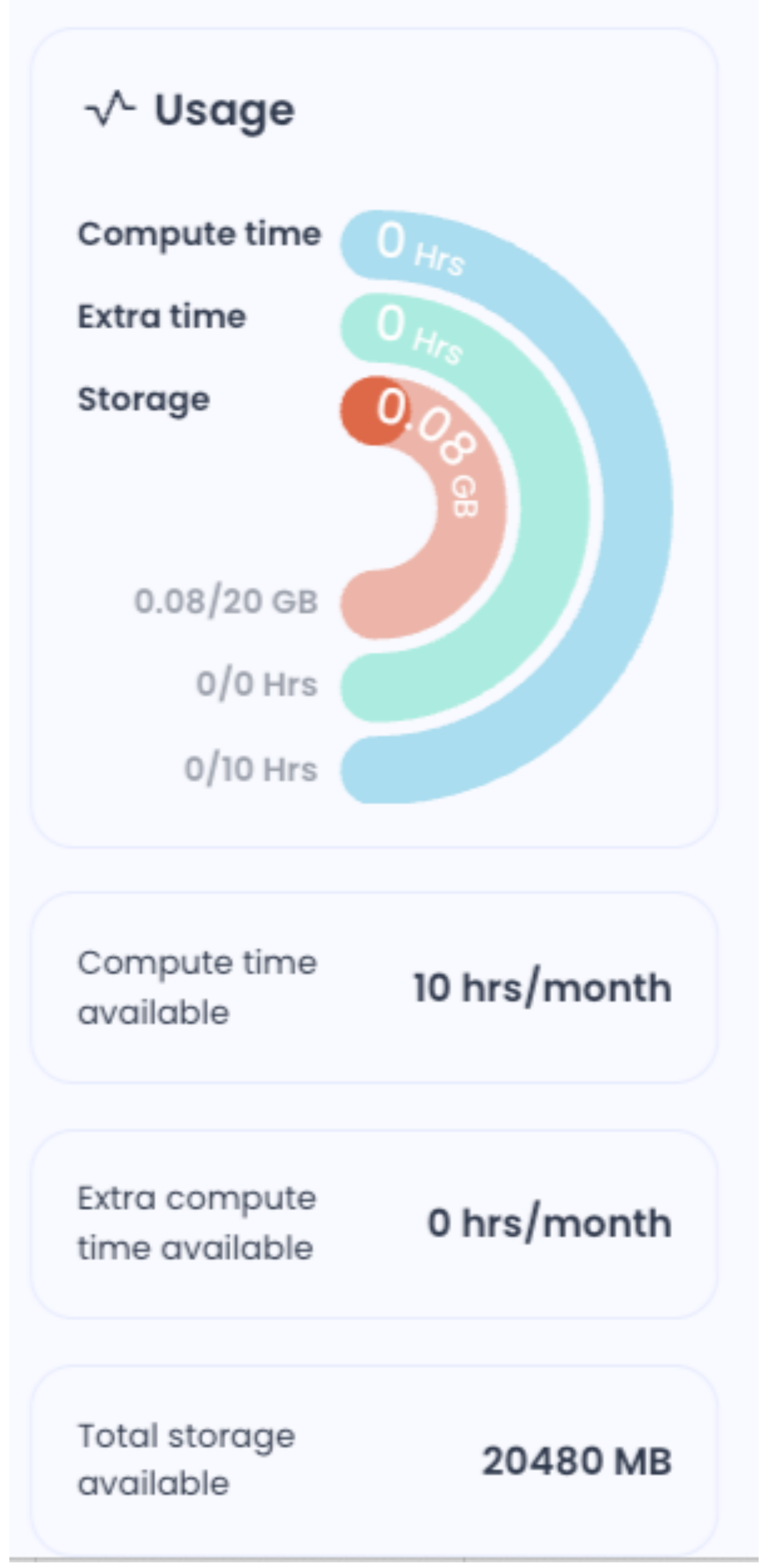
MATLAB for Open Science?

At least they also know it "sells" for now



CodeOcean: Open Science Library

It's free for now... but why not institute-level license?



Usage

Compute time

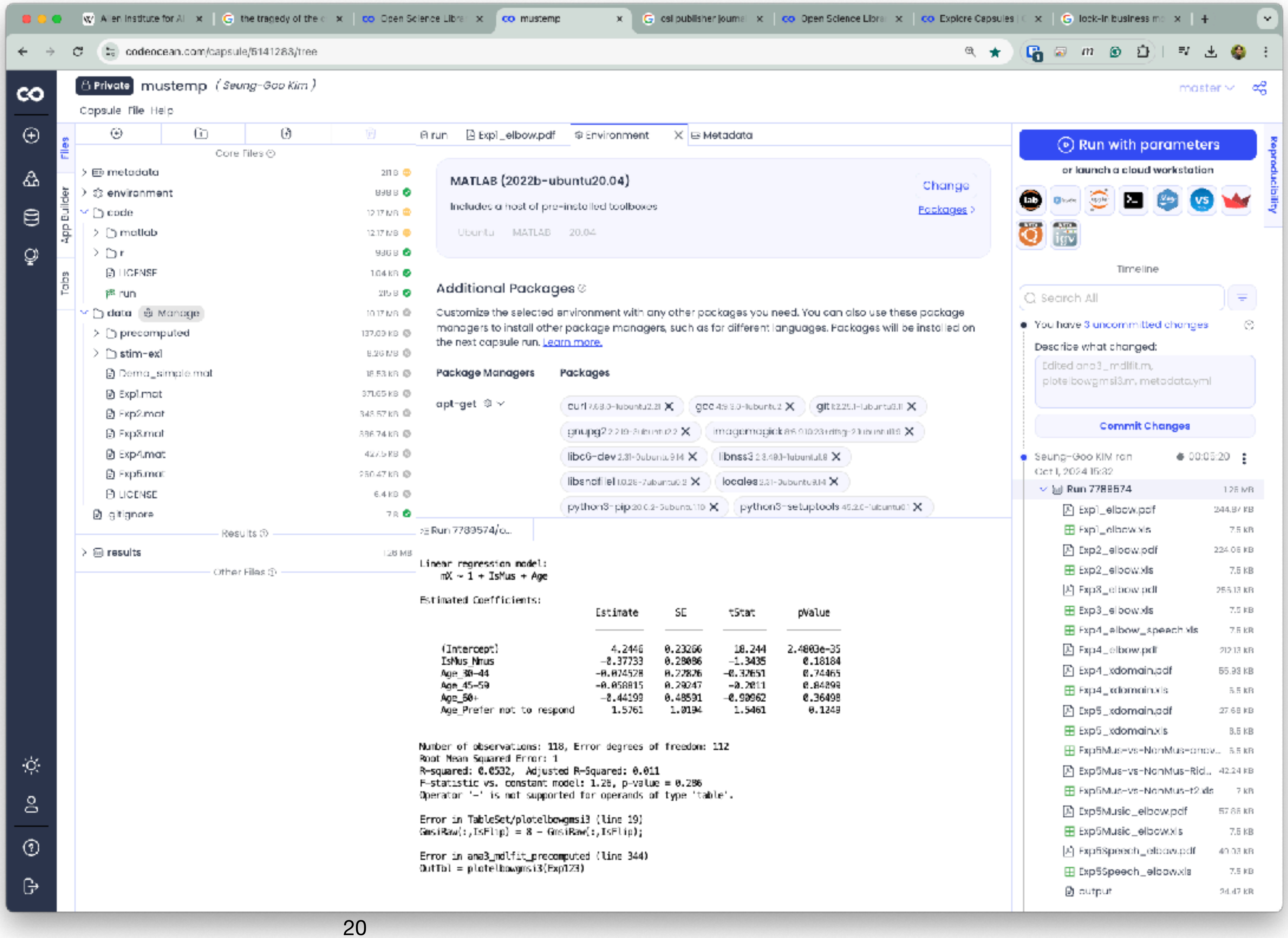
Extra time

Storage

Compute time available 10 hrs/month

Extra compute time available 0 hrs/month

Total storage available 20480 MB



MATLAB (2022b-ubuntu20.04)

Includes a host of pre-installed toolboxes

Ubuntu MATLAB 20.04

Change

Packages >

Additional Packages

Customize the selected environment with any other packages you need. You can also use these package managers to install other package managers, such as for different languages. Packages will be installed on the next capsule run. Learn more.

Package Managers	Packages
apt-get	curl, gcc, git, gnupg2, imagemagick, libc6-dev, libnss3, libsndfile1, locales, python3-pip, python3-setuptools

Run 7789574/o...

```
Linear regression model:
    mX ~ 1 + IsMus + Age

Estimated Coefficients:
                                 Estimate       SE        tStat       pValue
                                 ________    _______    ________    __________

    (Intercept)                    4.2446    0.23266      18.244    2.4803e-35
    IsMus_Nmus                   -0.37733    0.28086     -1.3435       0.18184
    Age_30-44                   -0.074528    0.22826    -0.32651       0.74465
    Age_45-59                   -0.058815    0.29247     -0.2011       0.84098
    Age_60+                      -0.44199    0.48591    -0.90962       0.36498
    Age_Prefer not to respond      1.5761     1.0194      1.5461        0.1249


Number of observations: 118, Error degrees of freedom: 112
Root Mean Squared Error: 1
R-squared: 0.0532,  Adjusted R-Squared: 0.011
F-statistic vs. constant model: 1.26, p-value = 0.286
Operator '-' is not supported for operands of type 'table'.

Error in TableSet/plotelbowgmsi3 (line 19)
GmsiRaw(:,IsFlip) = 8 - GmsiRaw(:,IsFlip);

Error in ana3_mdlfit_precomputed (line 344)
OutTbl = plotelbowgmsi3(Exp123)
```

You have 3 uncommitted changes

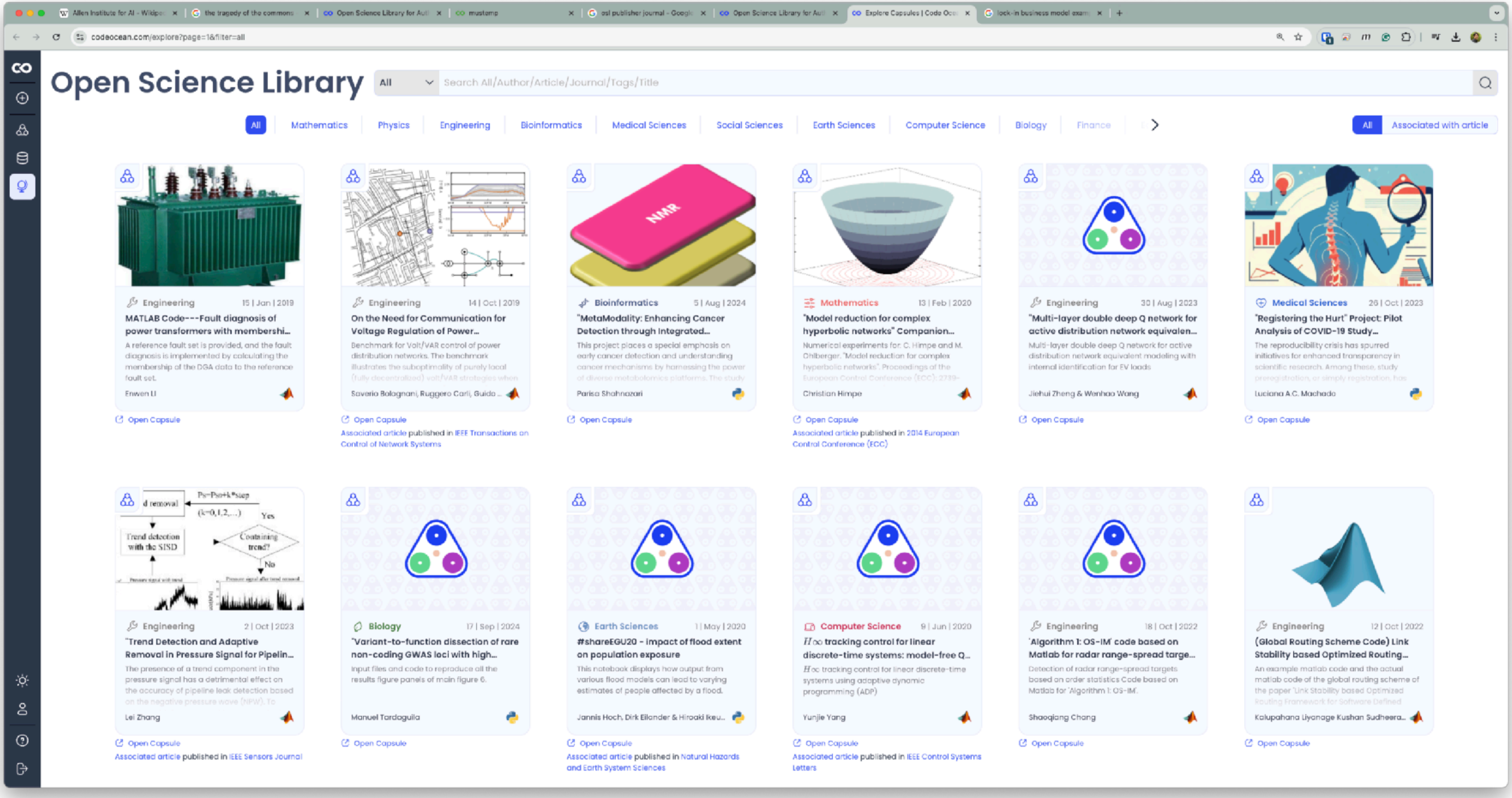
Describe what changed:

Edited ana3_mdlfit.m, plotelbowgmsi3.m, metadata.yml

Commit Changes

Seung-Goo KIM ran 00:05:20
Oct 1, 2024 15:32

Run 7788574 128 MB


Allen Institute for AI - Wikipe
the tragedy of the commons
Open Science Library for Aut
mustemp
osl publisher journal - Google
Open Science Library for Aut
Explore Capsules | Code Oce
lock-in business model exam
codeocean.com/explore?page=1&filter=all
All
Search All/Author/Article/Journal/Tags/Title
NMR
Ps=Ps0+k*step
(k=0,1,2,...)
Yes
No
Trend detection with the SISD
Containing trend?

Open Science Library

All | Mathematics | Physics | Engineering | Bioinformatics | Medical Sciences | Social Sciences | Earth Sciences | Computer Science | Biology | Finance

All | Associated with article

Engineering 15 | Jan | 2019
MATLAB Code---Fault diagnosis of power transformers with membershi...
A reference fault set is provided, and the fault diagnosis is implemented by calculating the membership of the DGA data to the reference fault set.
Enwen Li
Open Capsule

Engineering 14 | Oct | 2019
On the Need for Communication for Voltage Regulation of Power...
Benchmark for Volt/VAR control of power distribution networks. The benchmark illustrates the suboptimality of purely local (fully decentralized) volt/VAR strategies when
Saverio Bolognani, Ruggero Carli, Guido ...
Open Capsule
Associated article published in IEEE Transactions on Control of Network Systems

Bioinformatics 5 | Aug | 2024
"MetaModality: Enhancing Cancer Detection through Integrated...
This project places a special emphasis on early cancer detection and understanding cancer mechanisms by harnessing the power of diverse metabolomics platforms. The study
Parisa Shahnazari
Open Capsule

Mathematics 13 | Feb | 2020
"Model reduction for complex hyperbolic networks" Companion...
Numerical experiments for: C. Himpe and M. Ohlberger. "Model reduction for complex hyperbolic networks". Proceedings of the European Control Conference (ECC): 2739-
Christian Himpe
Open Capsule
Associated article published in 2014 European Control Conference (ECC)

Engineering 30 | Aug | 2023
"Multi-layer double deep Q network for active distribution network equivalen...
Multi-layer double deep Q network for active distribution network equivalent modeling with internal identification for EV loads
Jiehui Zheng & Wenhao Wang
Open Capsule

Medical Sciences 26 | Oct | 2023
"Registering the Hurt" Project: Pilot Analysis of COVID-19 Study...
The reproducibility crisis has spurred initiatives for enhanced transparency in scientific research. Among these, study preregistration, or simply registration, has
Luciana A.C. Machado
Open Capsule

Engineering 2 | Oct | 2023
"Trend Detection and Adaptive Removal in Pressure Signal for Pipelin...
The presence of a trend component in the pressure signal has a detrimental effect on the accuracy of pipeline leak detection based on the negative pressure wave (NPW). To
Lei Zhang
Open Capsule
Associated article published in IEEE Sensors Journal

Biology 17 | Sep | 2024
"Variant-to-function dissection of rare non-coding GWAS loci with high...
Input files and code to reproduce all the results figure panels of main figure 6.
Manuel Tardaguila
Open Capsule

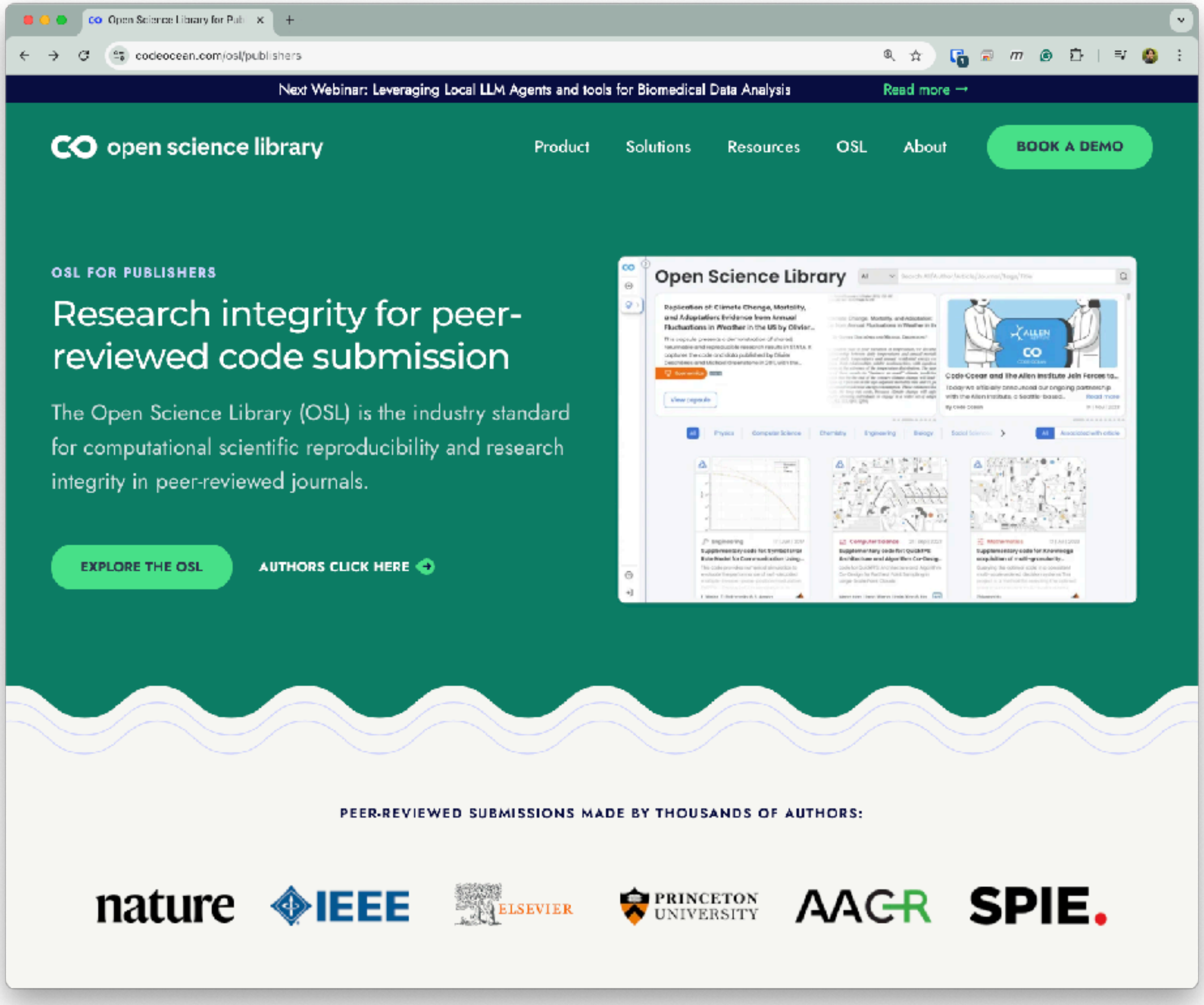
Earth Sciences 1 | May | 2020
#shareEGU20 - impact of flood extent on population exposure
This notebook displays how output from various flood models can lead to varying estimates of people affected by a flood.
Jannis Hoch, Dirk Eilander & Hiroaki Ikeu...
Open Capsule
Associated article published in Natural Hazards and Earth System Sciences

Computer Science 9 | Jun | 2020
$H\infty$ tracking control for linear discrete-time systems: model-free Q...
$H\infty$ tracking control for linear discrete-time systems using adaptive dynamic programming (ADP)
Yunjie Yang
Open Capsule
Associated article published in IEEE Control Systems Letters

Engineering 18 | Oct | 2022
'Algorithm 1: OS-IM' code based on Matlab for radar range-spread targe...
Detection of radar range-spread targets based on order statistics Code based on Matlab for 'Algorithm 1: OS-IM'.
Shaoqiang Chang
Open Capsule

Engineering 12 | Oct | 2022
(Global Routing Scheme Code) Link Stability based Optimized Routing...
An example matlab code and the actual matlab code of the global routing scheme of the paper 'Link Stability based Optimized Routing Framework for Software Defined
Kalupahana Liyanage Kushan Sudheera...
Open Capsule

For journals



codeocean.com/osl/publishers

Next Webinar: Leveraging Local LLM Agents and tools for Biomedical Data Analysis Read more →

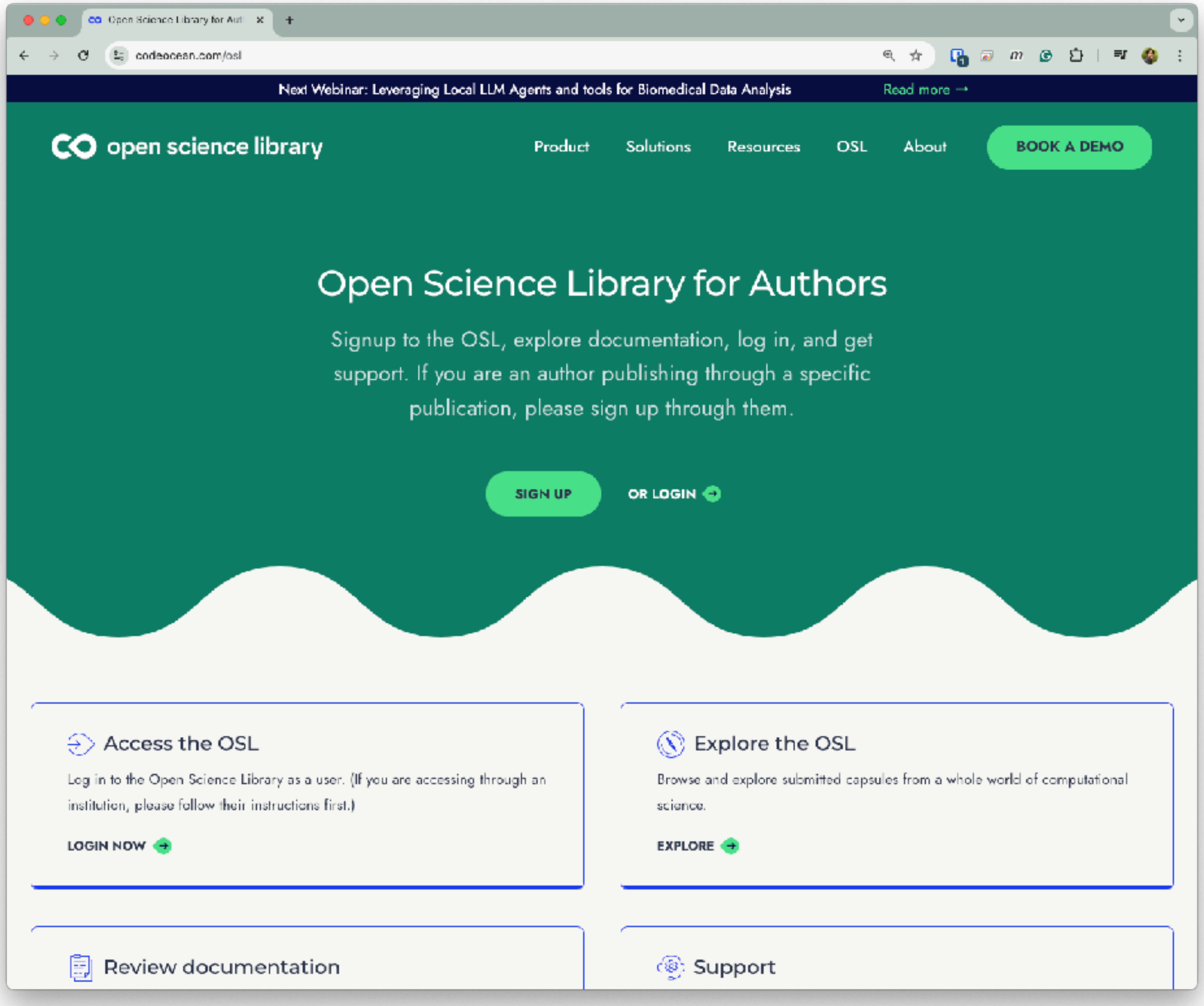
OSL FOR PUBLISHERS

Research integrity for peer-reviewed code submission

The Open Science Library (OSL) is the industry standard for computational scientific reproducibility and research integrity in peer-reviewed journals.

PEER-REVIEWED SUBMISSIONS MADE BY THOUSANDS OF AUTHORS:

For authors



codeocean.com/osl

Next Webinar: Leveraging Local LLM Agents and tools for Biomedical Data Analysis Read more →

Open Science Library for Authors

Signup to the OSL, explore documentation, log in, and get support. If you are an author publishing through a specific publication, please sign up through them.

Access the OSL

Log in to the Open Science Library as a user. (If you are accessing through an institution, please follow their instructions first.)

Explore the OSL

Browse and explore submitted capsules from a whole world of computational science.

Review documentation

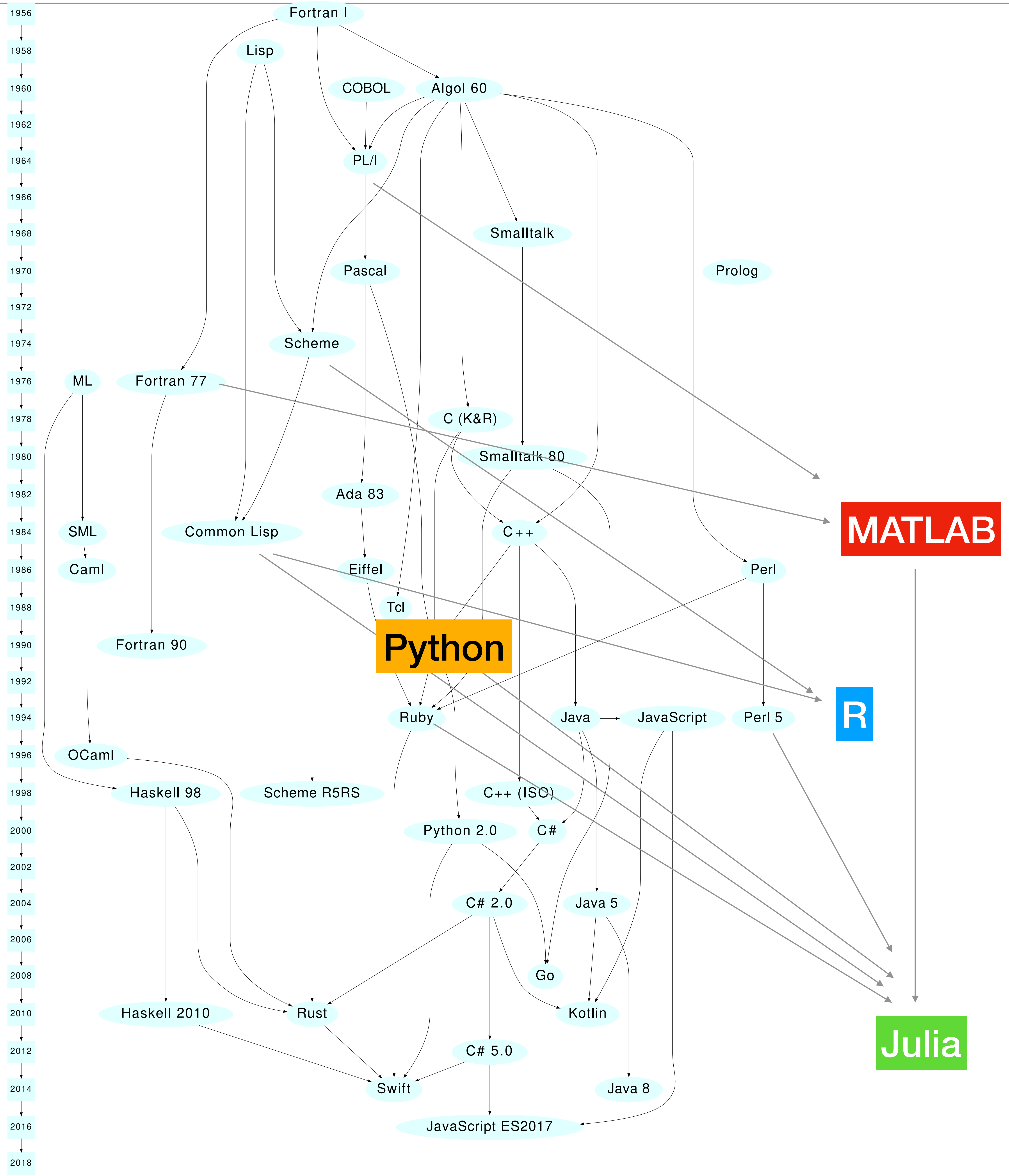
Support

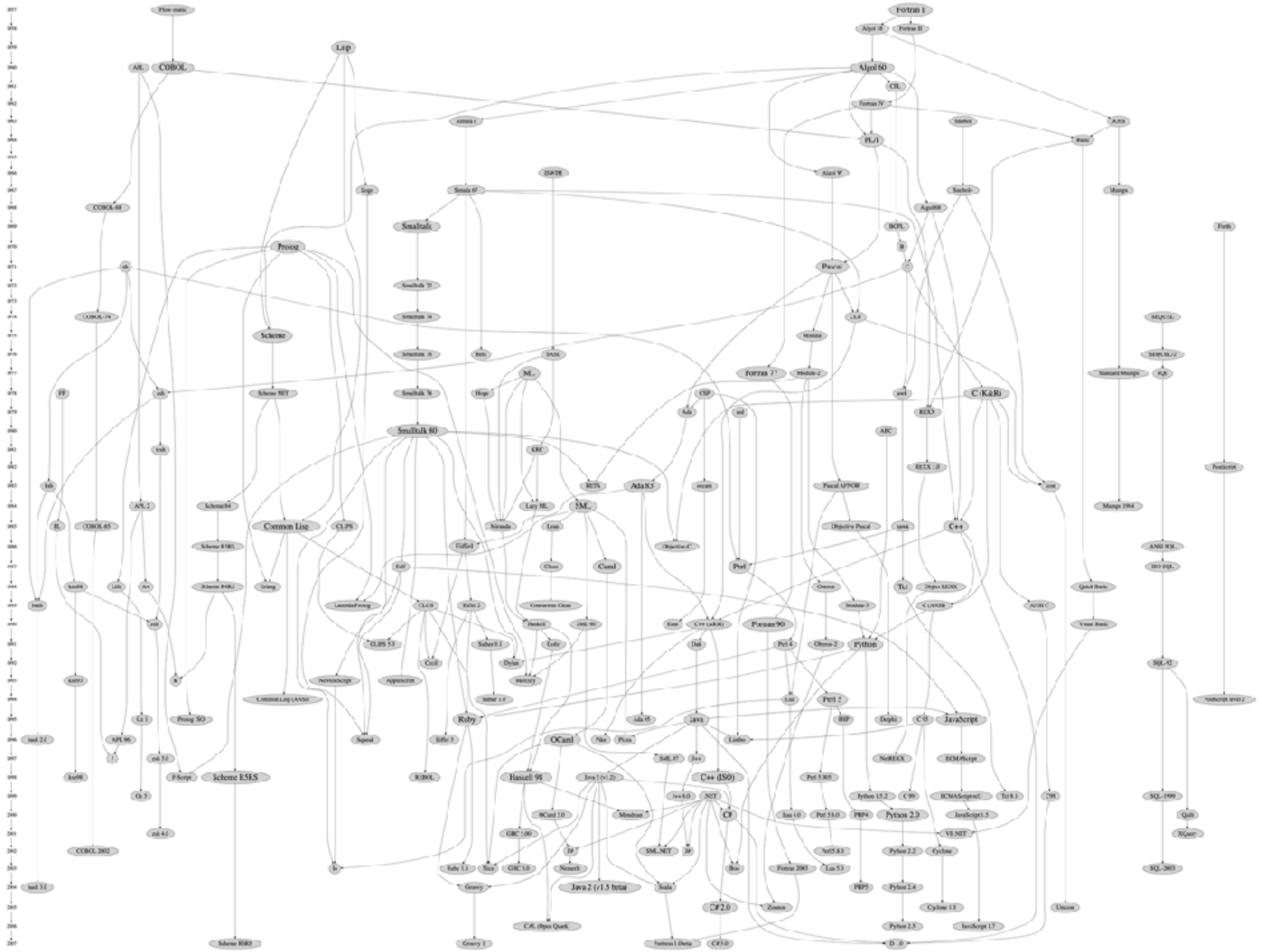
Your thoughts? 🧐💭

Open questions

- Is proprietary software evil? (should/can we be non-evil in the long run?)
- Can the transparency be implemented only by open-access software?
- What is the best language?

Based on wikipedia [2024-10-01] https://github.com/stereobooster/programming-languages-genealogical-tree?tab=readme-ov-file





What is the best computing language?

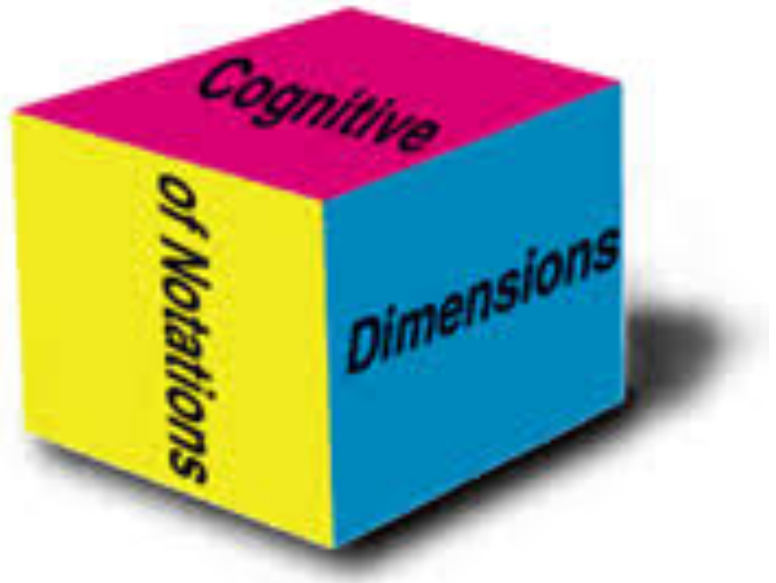
Computing languages as meta-tools to build your own tools

- As an end-user of a language, we want tools with high...
  - usability [HCI]: easy to learn/read/write/debug/maintain
  - accessibility [OS]: can be used without paid licenses
  - sustainability [OS, Econ]: 50+ years
  - performance [CS]: can handle large data
- ... in a good balance.

Language	Latest release	GPL/DSL	Usability	Accessibility	Sustainability	Performance
Python	2024-10-01	GPL	?	(600$/yr/user?)	?	mid
GNU-c-LISP	2023-01-13	GPL	?	0	?	?
C	2024-02-21	GPL	?	0	?	high
C++	2023-03-19	GPL	?	0	?	high
C# (Win)	2023-11-14	GPL	?	0	?	high
Swift (Mac)	2024-03-01	GPL	?	0	?	high
Julia	2024-10-01	GPL	?	0	?	mid
Java	2024-09-17	GPL	?	180$/yr/user	?	high
R	2024-06-14	DSL	?	0	?	?
GNU Octave	2024-06-07	DSL	?	0	?	mid
Fortran	2023-11-17	DSL	?	0	?	high
Mathematica	2024-07-01	DSL	?	839$/yr/user	?	?
Stata	2023-04-25	DSL	?	925$/yr/user	?	?
WolframOne	2024-07-31	DSL	?	1710$/yr/user	?	?
MATLAB	2024-09-12	DSL	?	263-3k$/yr/user	?	mid

HCI research

ChatGPT4o says...



- The **Cognitive Dimensions of Notations (CDN)** is a widely-used framework for evaluating the usability of notational systems, including DSLs (domain-specific languages). It provides a structured way to analyze how the design of a DSL affects the cognitive load on users. Some key dimensions include:
  - **Closeness of Mapping**: How well the language matches the mental model of the domain.
  - **Consistency**: How predictable and uniform the language is.
  - **Abstraction Gradient**: The complexity and flexibility in expressing concepts.
  - **Error-proneness**: How likely users are to make mistakes.
  - **Viscosity**: Resistance to change or how hard it is to make modifications.

Closeness of Mapping:
Let's invert a matrix !

<Generated by ChatGPT4o>

DEMO EXAMPLES

$$\mathbf{A}^+, (a_{i,j}) \in \mathbb{R}^{n\times m}, n > m$$

```
<Python>
import numpy as np
A = np.random.rand(10, 5)
A_pinv = np.linalg.pinv(A)
```

```
<R>
A <- matrix(runif(50), 10, 5)
A_pinv <- MASS::ginv(A)
```

```
<octave; MATLAB; Julia>
A = rand(10, 5)
A_pinv = pinv(A)
```

```
<APL>
A ← 10 5 ρ ? 100
+⌹A
```

```
<Perl+PDL>
use PDL;
$A = random 10,5;
$A_pinv = pinv($A);
```

```
<Cpp>
#include <iostream>
#include <armadillo>
using namespace arma;
int main() {
   mat A = randu<mat>(10, 5);
   mat A_pinv = pinv(A);
```

```
<Fortran>
program pseudoinverse
  implicit none
  integer, parameter :: m = 10, n = 5
  real(8), dimension(m,n) :: A
  real(8), dimension(n,n) :: A_pinv
  real(8), dimension(m) :: singular_values
  real(8), dimension(m,m) :: U
  real(8), dimension(n,n) :: VT
  real(8), dimension(n) :: S_inv
  integer :: info, i, j
  real(8), external :: random_number

  call random_seed()
  do i = 1, m
```

```
<LaTex+SageMath>
\documentclass{article}
\usepackage{sagetex}
\begin{document}
\begin{sagesilent}
A = random_matrix(RDF, 10, 5)
A_pinv = A.pseudoinverse()
\end{sagesilent}
\[ \sage{A} \sage{A_[pinv} \]
\end{document}
```

```
<HTML+MathJS>
<!DOCTYPE html>
   <script src="https://cdnjs.cloudflare.com/
ajax/libs/mathjs/10.6.4/math.min.js"></
script>
<body>
   <script>
      let A = math.random([10, 5]);
      let A_pinv = math.pinv(A);
document.getElementById('output').innerHTML =
"Original Matrix (10x5):\n" + math.format(A,
{precision: 3}) + "\n\nPseudoinverse:\n" +
math.format(A_pinv, {precision: 3});
    </script>
</body>
```

My conclusions

- Still Python & R seem to be good options.
- Non-CS scientists love math-oriented languages like R, MATLAB, Julia (we still misunderstand that we are handling natural numbers)
- Julia looks very interesting (e.g., UTF-8 for all names: 🤨(a,b) = a^b; 파이=pi; 🤨(파이,3); built-in functions for math operations like MATLAB)
- MATLAB is so expensive... but so as many other tools we use (like Macs and Adobe suites).
- Usability is difficult to measure; and the field-leading developers decide which language we use.
- Transparency may be still achieved using specialized platforms like CodeOcean (it's not a trivial task that many individual researchers can handle)