Applying Foundational Open Science Skills (FOSS) in Public Health

Presented by

Roots for Resilience (R4R 2024 Fellows)

Imran Mithu

Caleigh Curley

Joy Luzingu

Overview of Root for Resilience (R4R)

RESEARCH, INNOVATION & IMPACT
Data Science Institute

CYVERSE

THE UNIVERSITY OF ARIZONA
Mel & Enid Zuckerman College of Public Health

What is R4R

Organizer

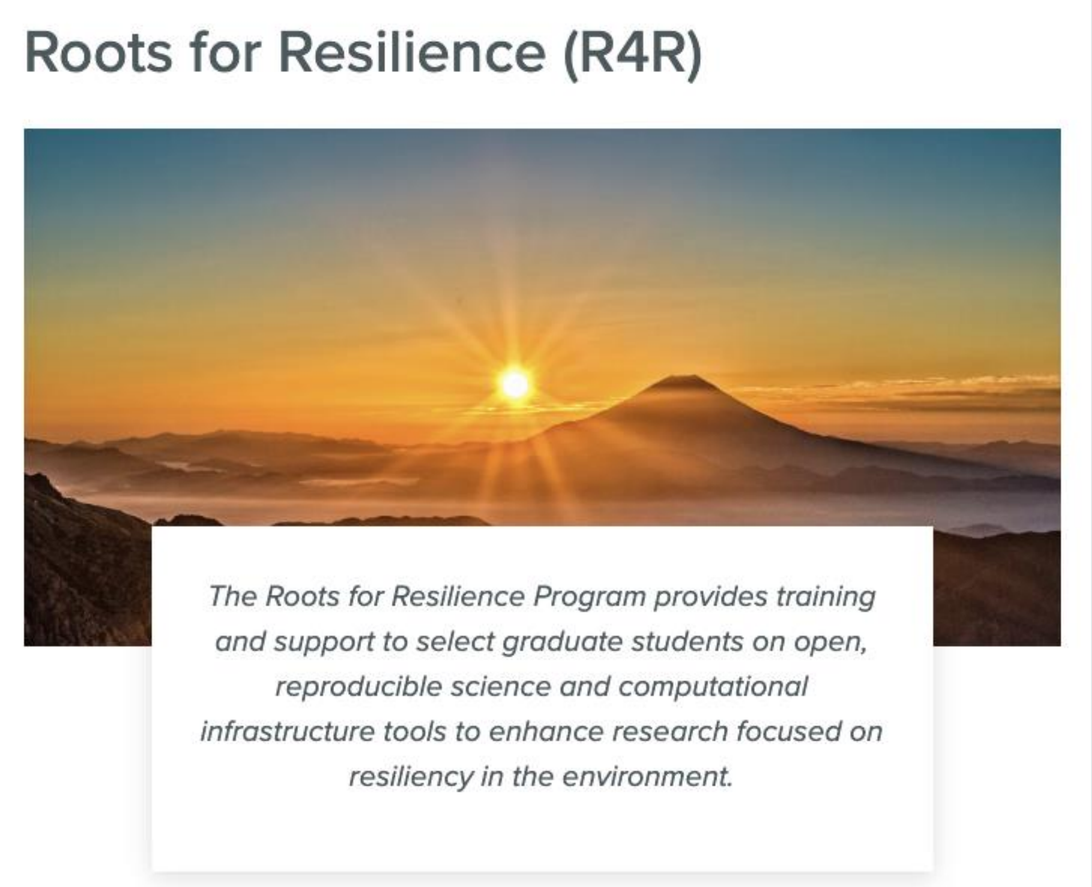
Arizona Institute for Resilience + Research, Innovation & Impact Data Science Institute + CyVerse

Who can participate?

Any graduate student of the U of A with the nomination of departmental head

Goal of R4R

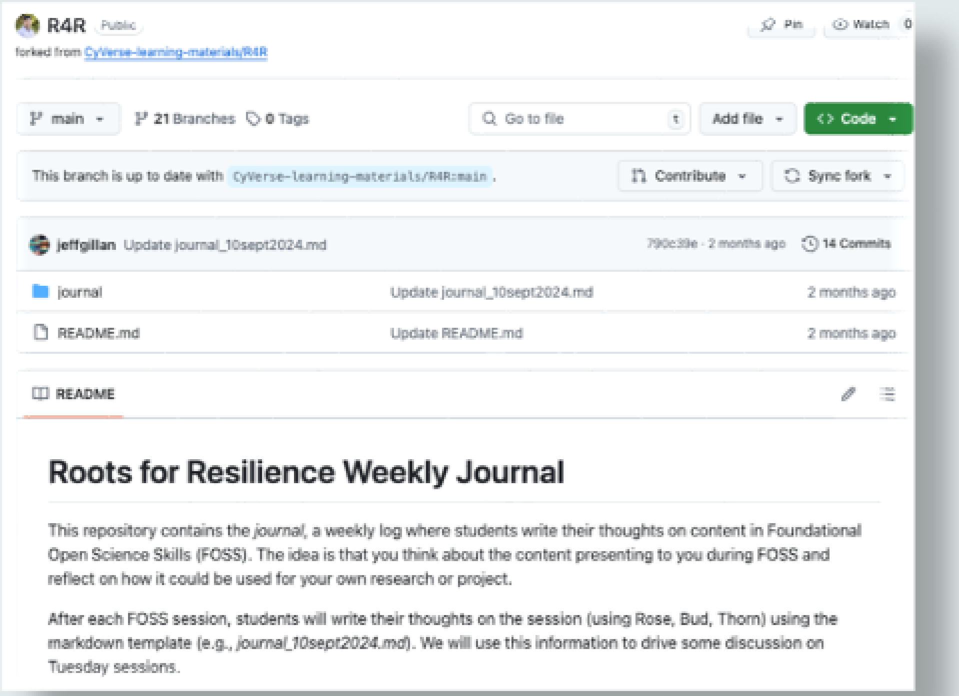
- Trains selected graduate students in the use of open science
- Apply data science tools to their dissertation research and discovery
- Increase their department's data science capacity



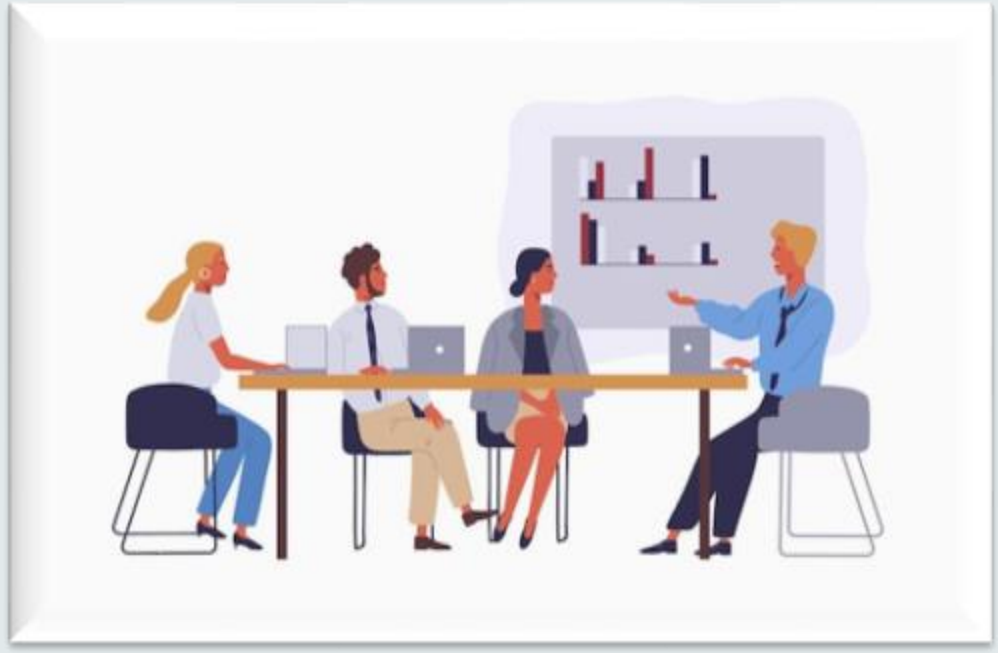
https://datascience.arizona.edu/r4r

What do we do during the fellowship?

Weekly Online FOSS Workshop on an Open Science Topic



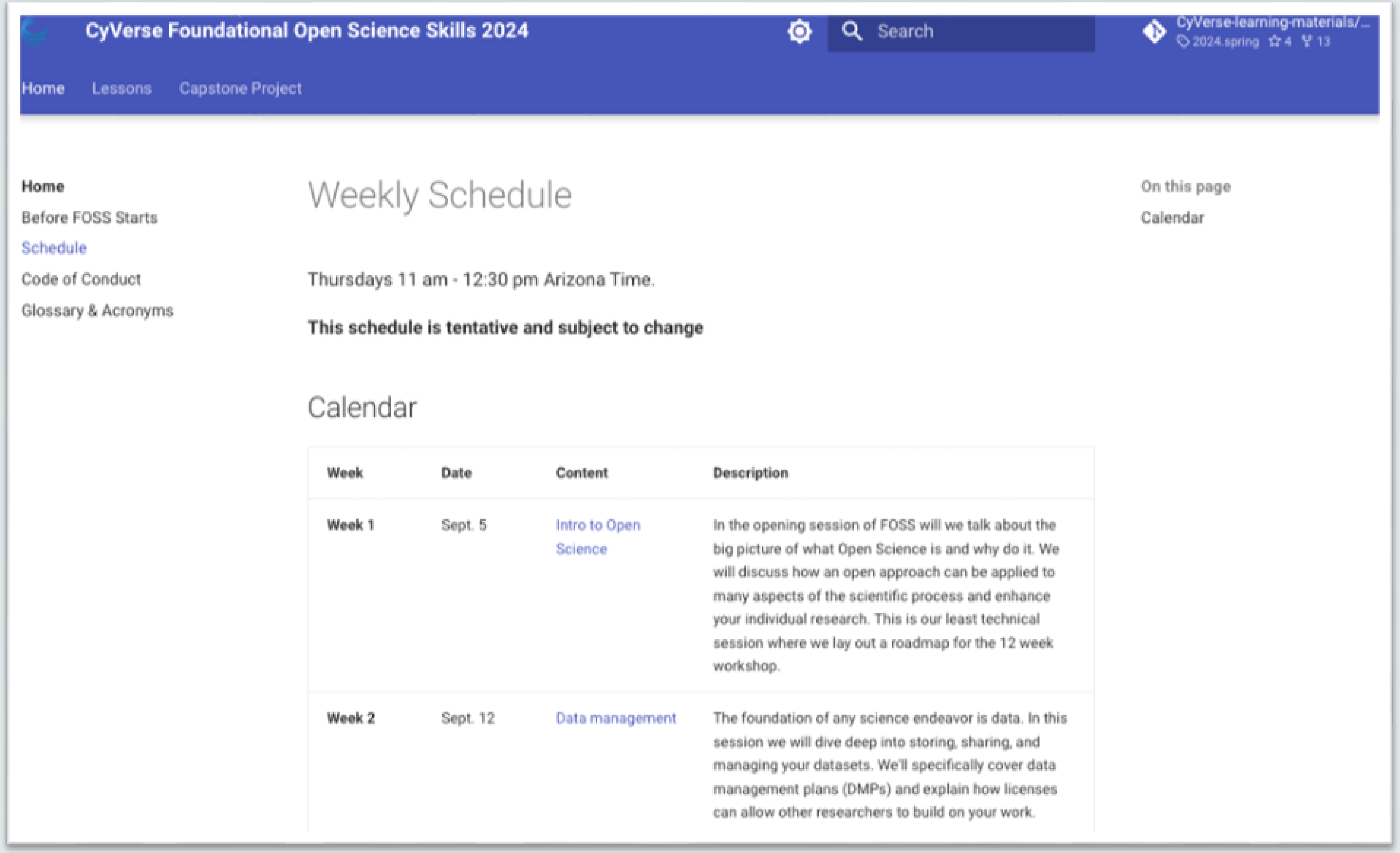
Weekly R4R Journal Submission



Weekly in-person R4R session

Content of the training

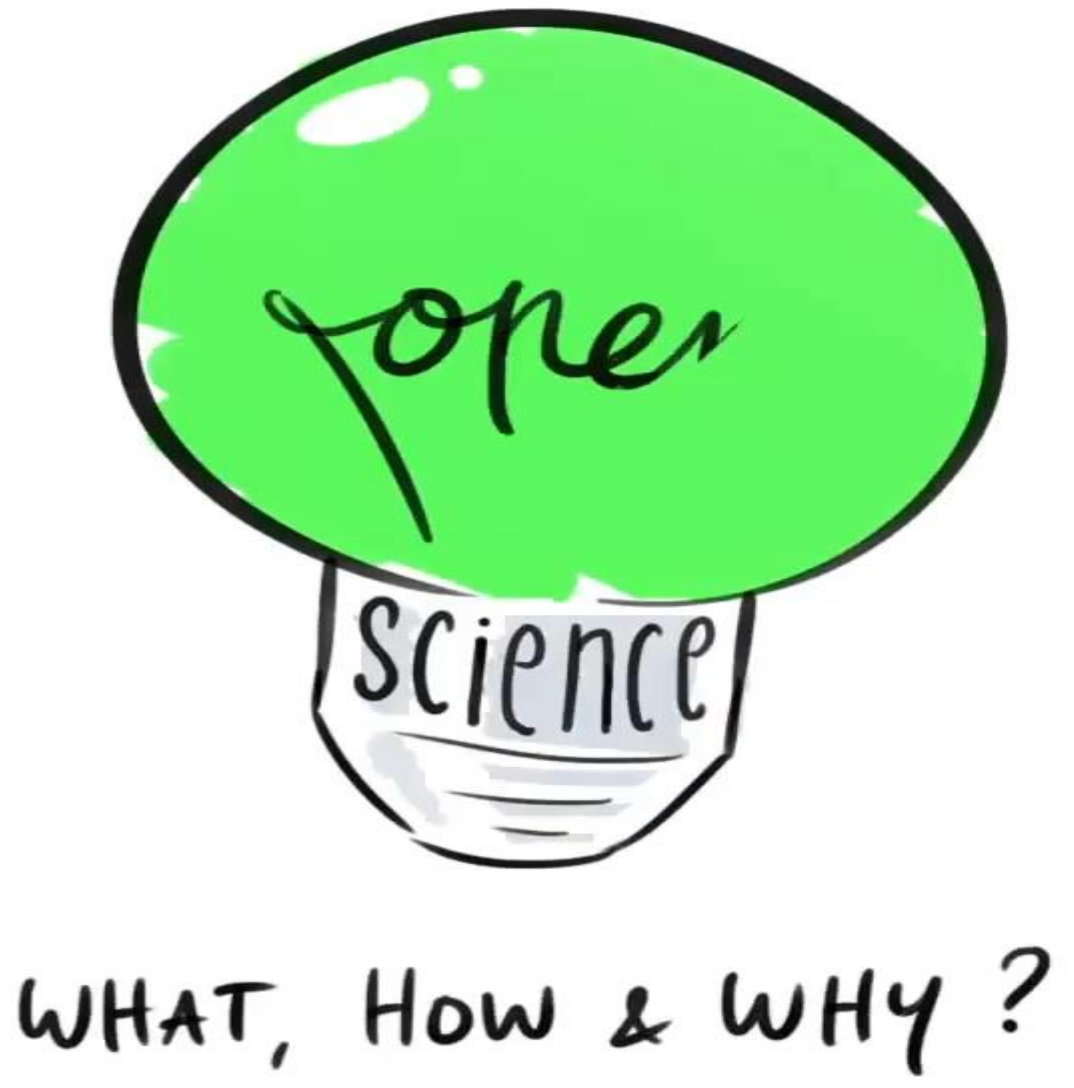
- Intro to Open Science
- Data Management
- Project Management
- Documentation and Communication
- Version Control
- Reproducibility
- Container Development
- High-Performance Computing (HPC)



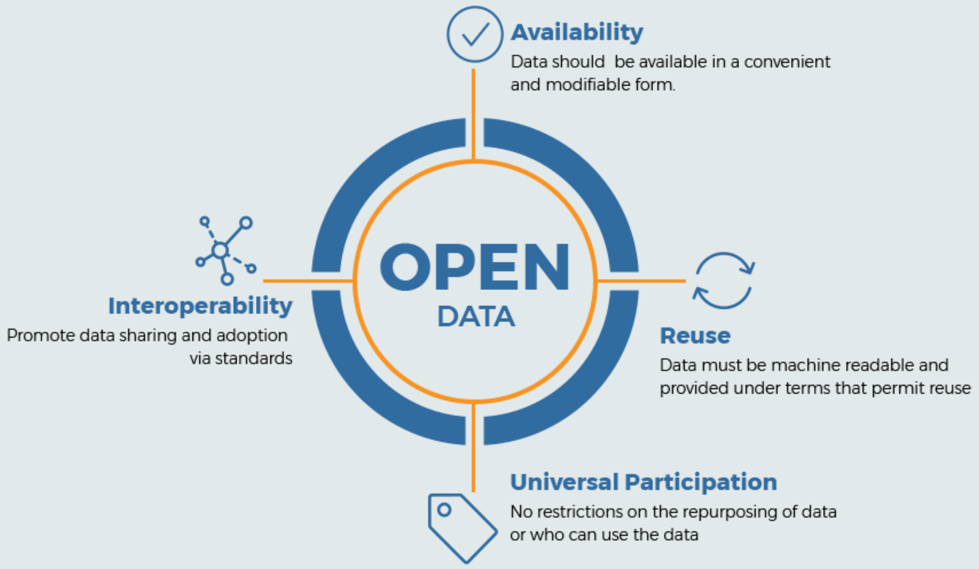
Website: https://foss.cyverse.org/schedule/

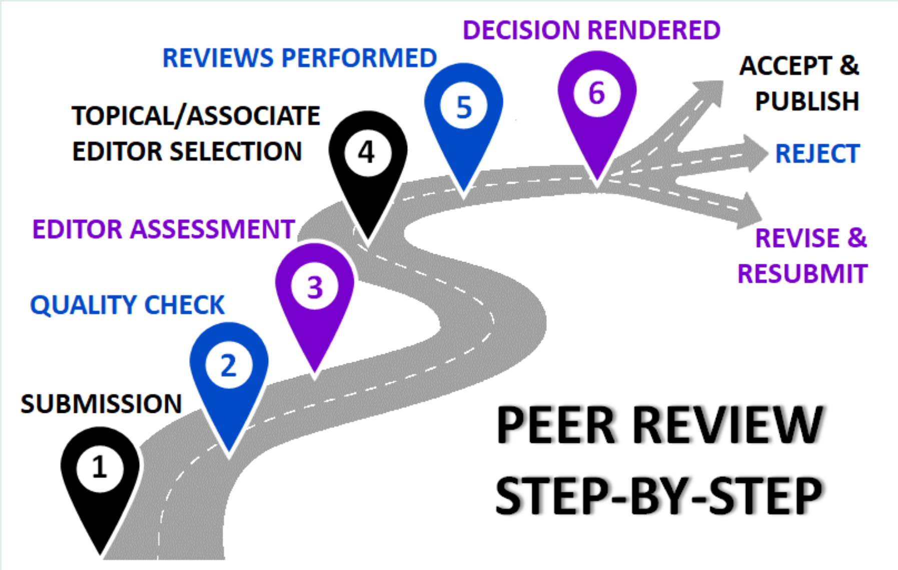
Overview of Open Science

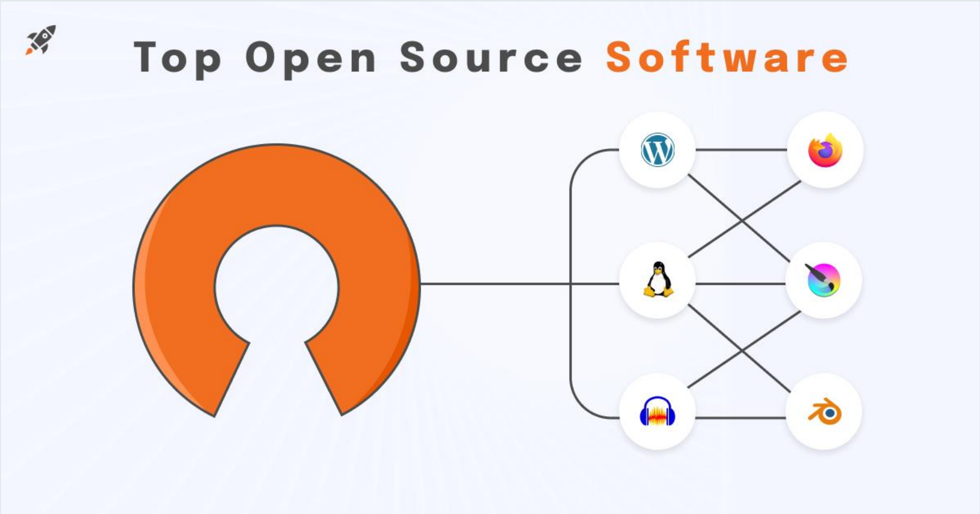
What is Open Science



Pillars of Open Science







https://foss.cyverse.org/01_intro_open_sci/

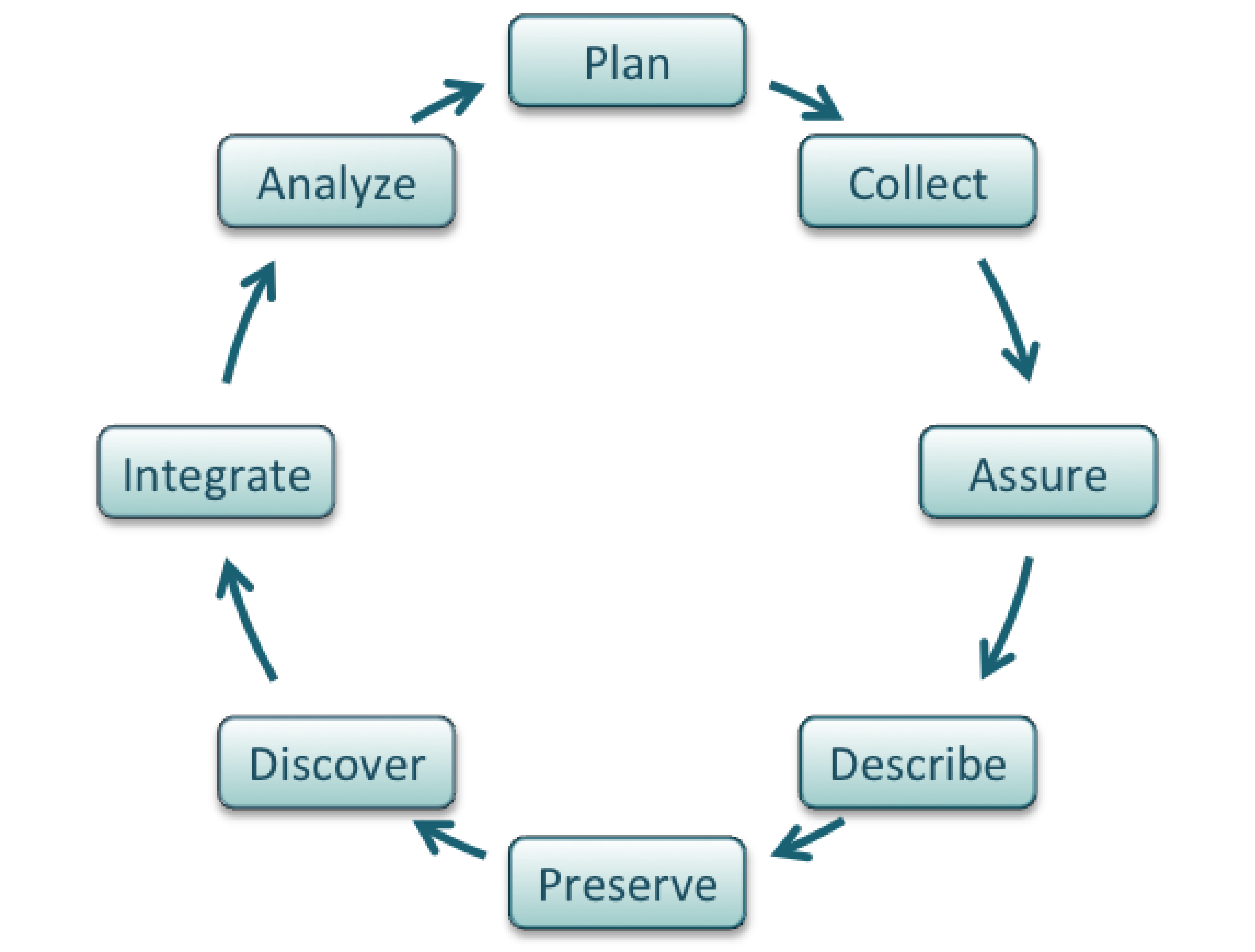
Open Science Framework

•

https://www.youtube.com/watch?v=X07mBq2tnMg&t=2s&ab_channel=CenterforOpenScience

Data Management

- Data management is the set of practices that allow researchers to effectively and efficiently handle data throughout the data life cycle.
- Although typically shown as a circle, the actual life cycle of any data item may follow a different path, with branches and internal loops.
- Being aware of your data's future helps you plan how to best manage them.



The Data Life Cycle, from Strasser et al.

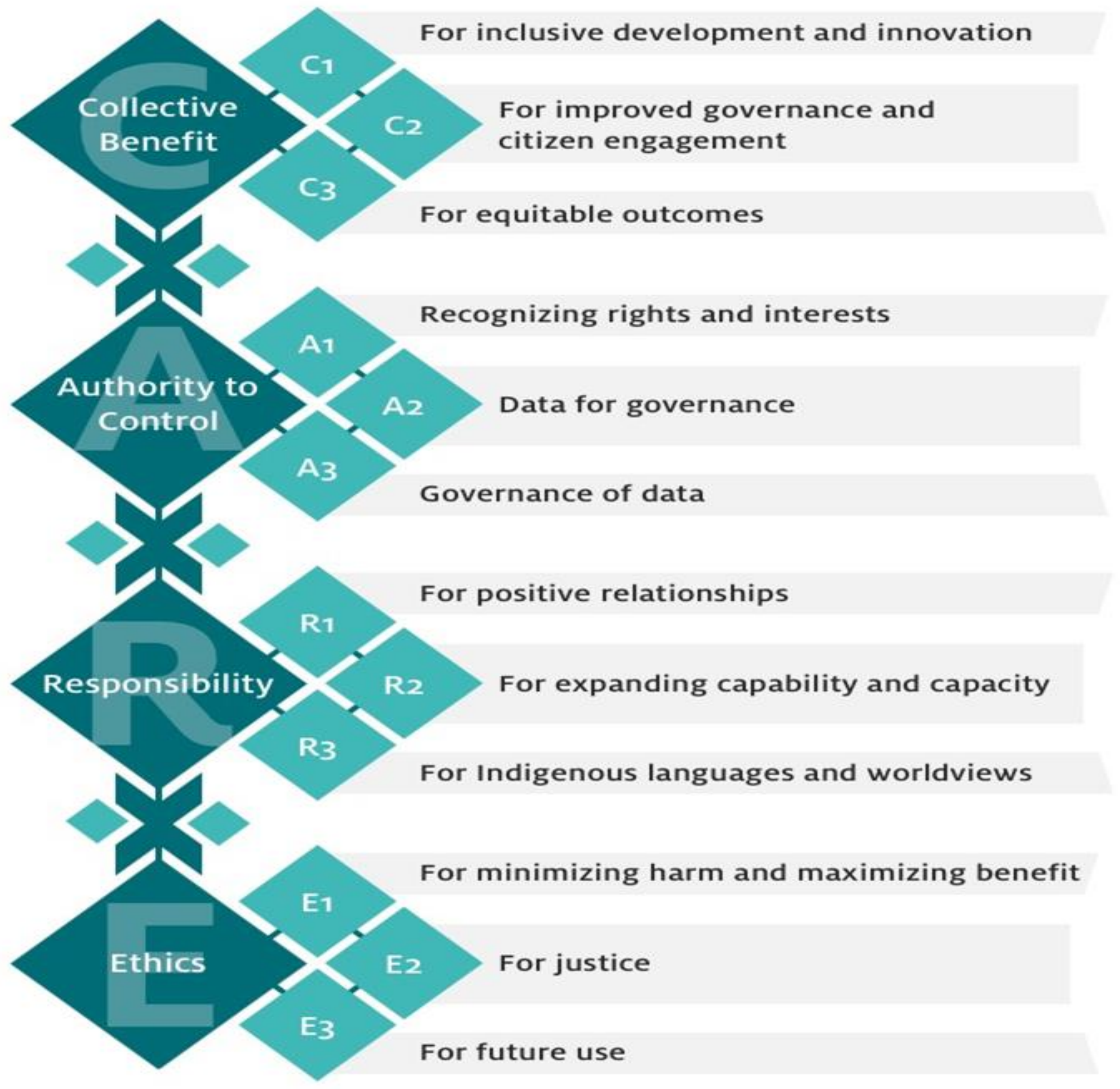
DATA PRINCIPLES

INDIGENOUS				MAINSTREAM		
New Zealand Indigenous Data Sovereignty Principles	**Australia Indigenous Data Sovereignty Protocols**	**United States Indigenous Data Governance Principles**	**Canada Indigenous Data Governance Principles**	**Open Data Charter Principles**	**FAIR Principles for Data Management and Stewardship**	**STREAM Properties for Industrial and Commoditized Data**
Authority	Self-Determination	Inherent Sovereignty	OCAP®	Open By Default	Findable	Sovereign
Relationships	Available and Accessible	Indigenous Knowledge	Indigenous Knowledge	Timely and Comprehensive	Accessible	Trusted
Obligations	Collective Rights and Interests	Ethics	Methodology and Approaches	Accessible and Usable	Interoperable	Reusable
Collective Benefit	Accountability	Intergenerational Collective Wellbeing	Evidence to Build Policy	Comparable and Interoperable	Reusable	Exchangeable
Reciprocity	Exercise Control	Relationships	Ethical Relationships	For Improved Governance & Citizen Engagement		Actionable
Guardianship			Data Governance	For Inclusive Development and Innovation		Measurable

People oriented principles | Purpose oriented principles | Data oriented principles

Carroll, S.R., Garba, I., Figueroa-Rodriguez, O.L., Holbrook, J., Lovett, R., Materrechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J., Hudson, M. 2020a. The CARE Principles for Indigenous Data Governance. Data Science Journal. 19 (43): 1-12.

Indigenous Frameworks



CARE Principles for Indigenous Data Governance

CARE Principles for Indigenous Data Governance

Collective Benefit.

Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.

C1. For inclusive development and innovation
C2. For improved governance and citizen engagement
C3. For equitable outcomes

Authority to Control.

Indigenous Peoples' rights and interests in Indigenous data must be recognized and their authority to control such data respected.

A1. Recognizing rights and interests
A2. Data for governance
A3. Governance of data

Responsibility.

Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous Peoples' self determination and collective benefit.

R1. For positive relationships
R2. For expanding capability and capacity
R3. For Indigenous languages and worldviews

Ethics.

Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

E1. For minimizing harm and maximizing benefit
E2. For justice
E3. For future use

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Be

FAIR

Findable Accessible Interoperable Reusable

and

CARE

Collective Benefit Authority to Control Responsibility Ethics

Documentation and Communication

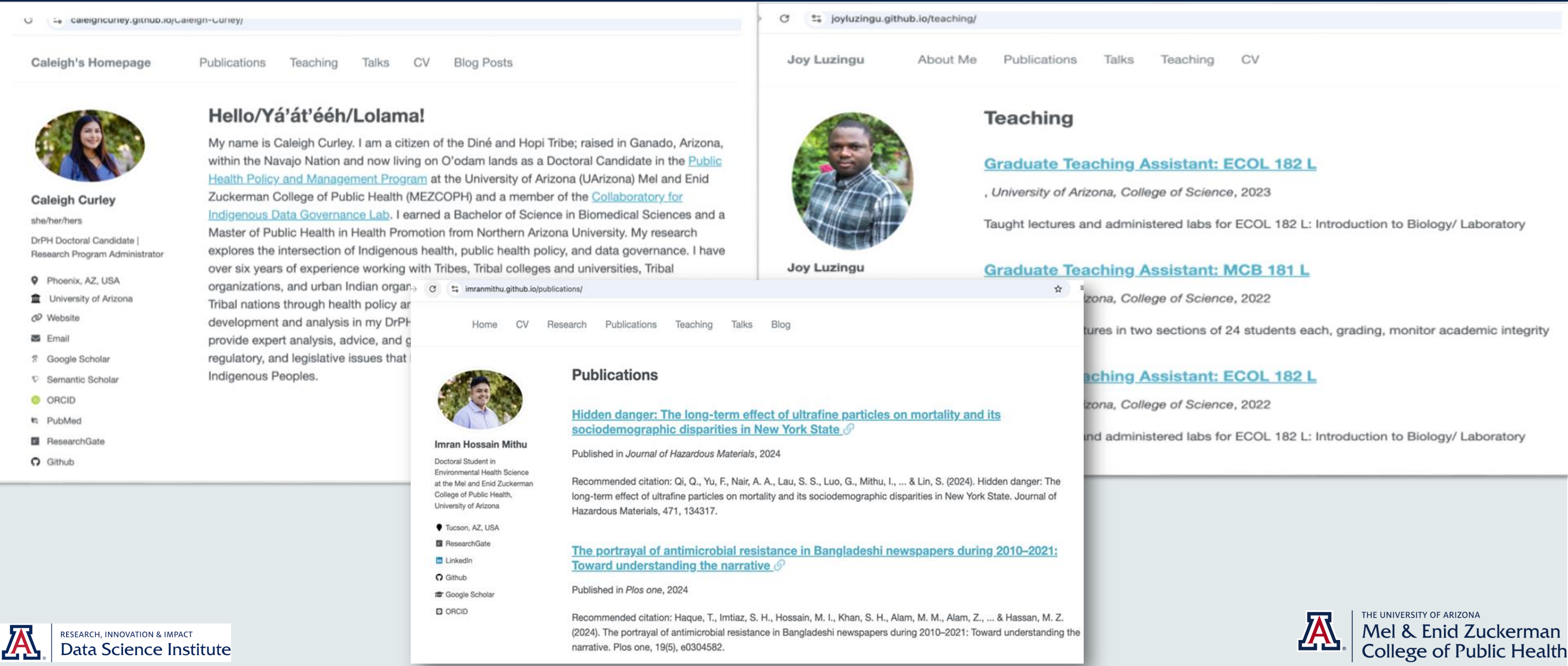
A great Open Scientist is someone who documents their work and shares it with the world. This means going well beyond peer-reviewed publications.

Public Repositories for Documentation

- GitHub Readme
- GitHub Wiki
- GitHub Pages
- Material MkDocs
- ReadTheDocs
- Bookdown
- Quarto
- JupyterBook
- GitBook
- Confluence Wikis

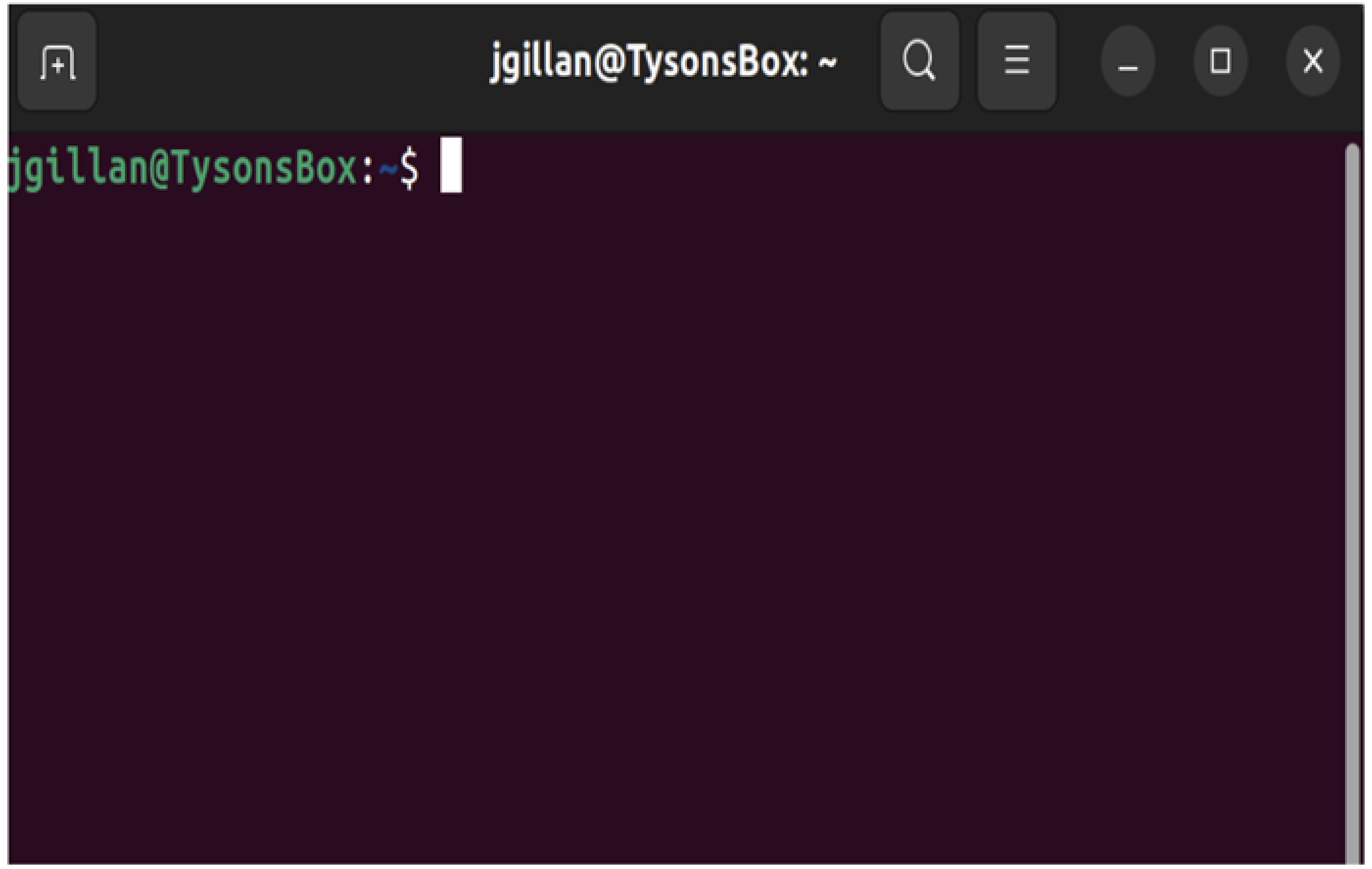
https://foss.cyverse.org/03_documentation_communication/#project-documentation

Documentation and Communication



How to talk to computers

The Command Line Interface (CLI)

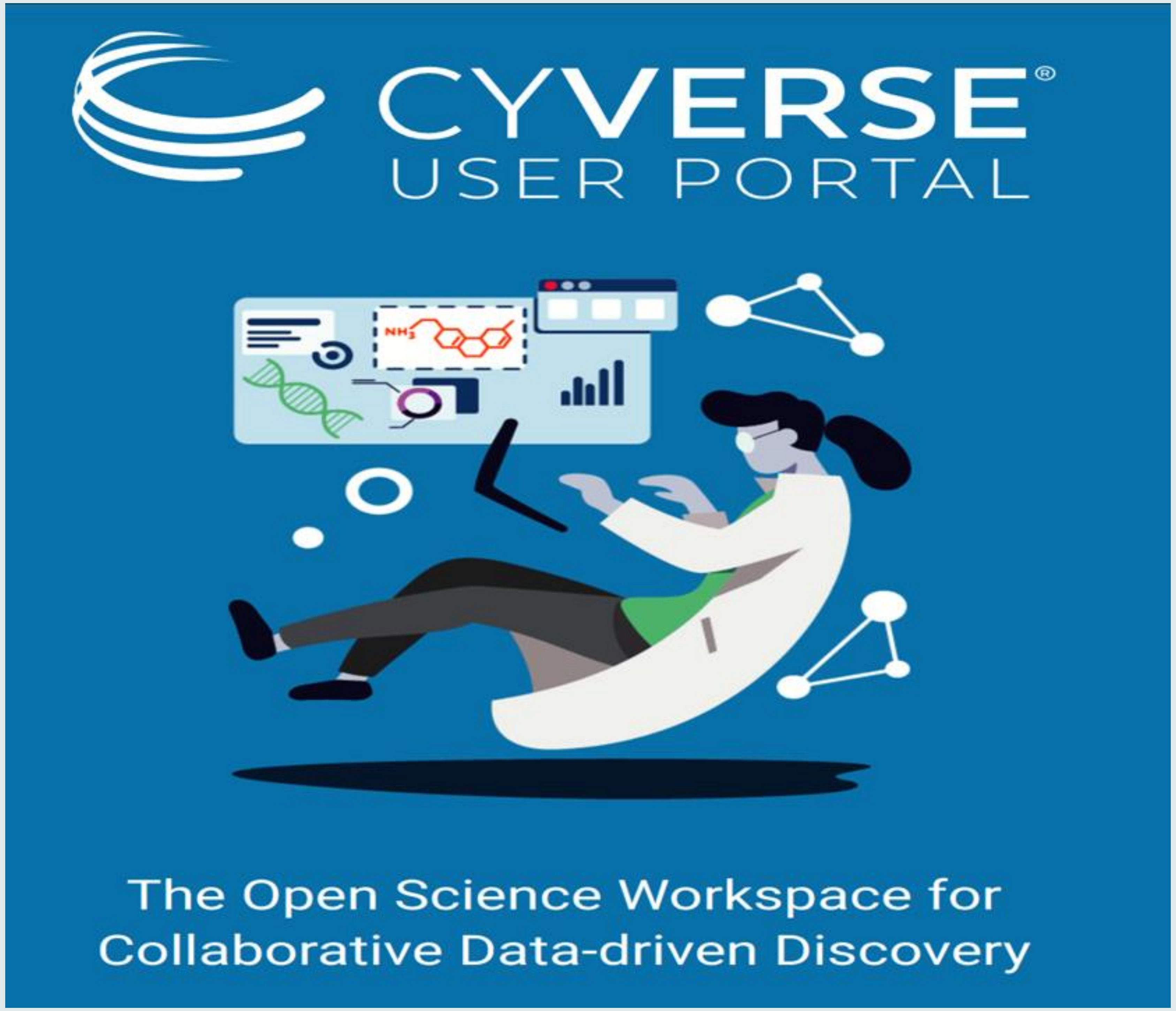


The Terminal shell

LLM Chatbots for Open Science

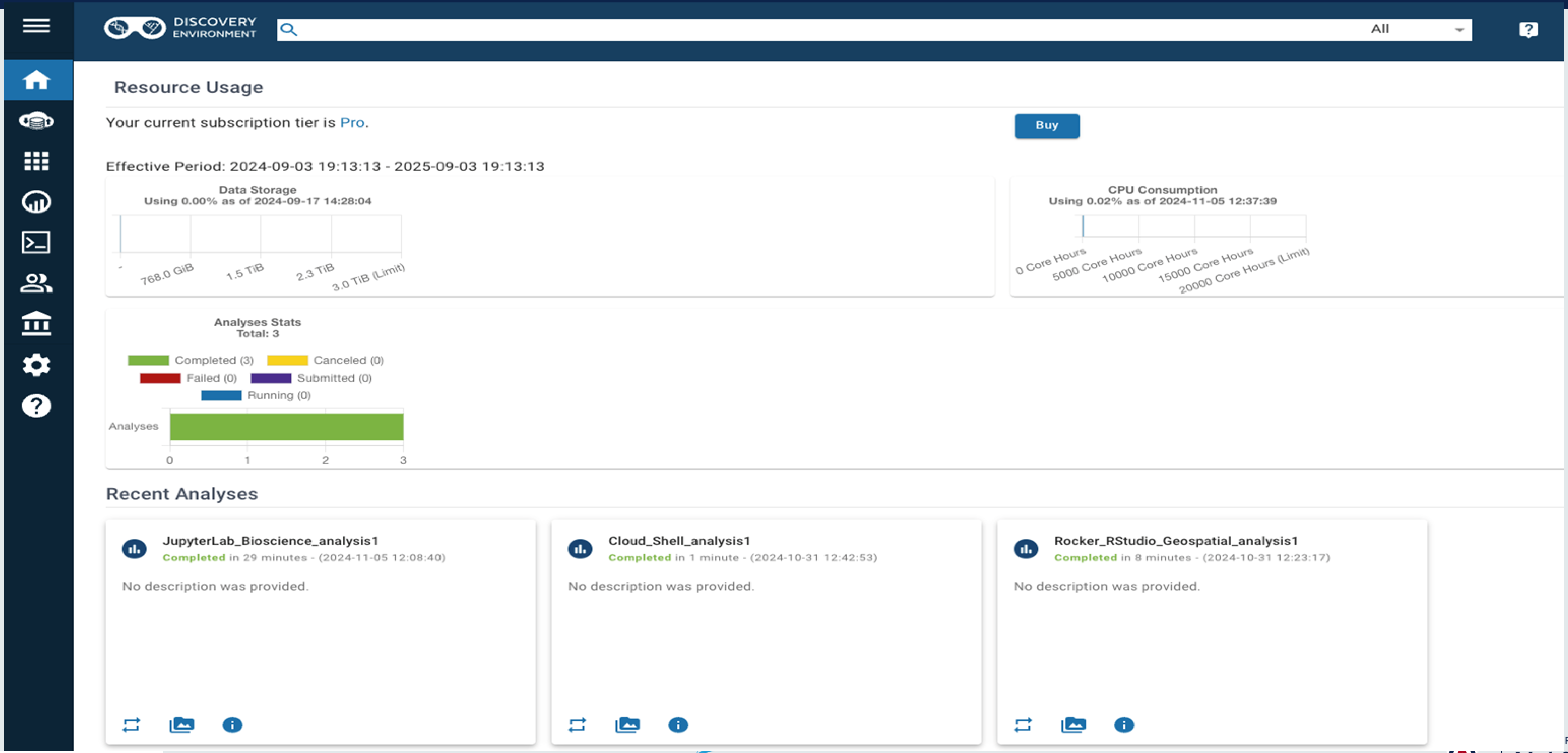
https://foss.cyverse.org/04_talk_to_computer/

Remote computing



https://cyverse.org/

Remote computing



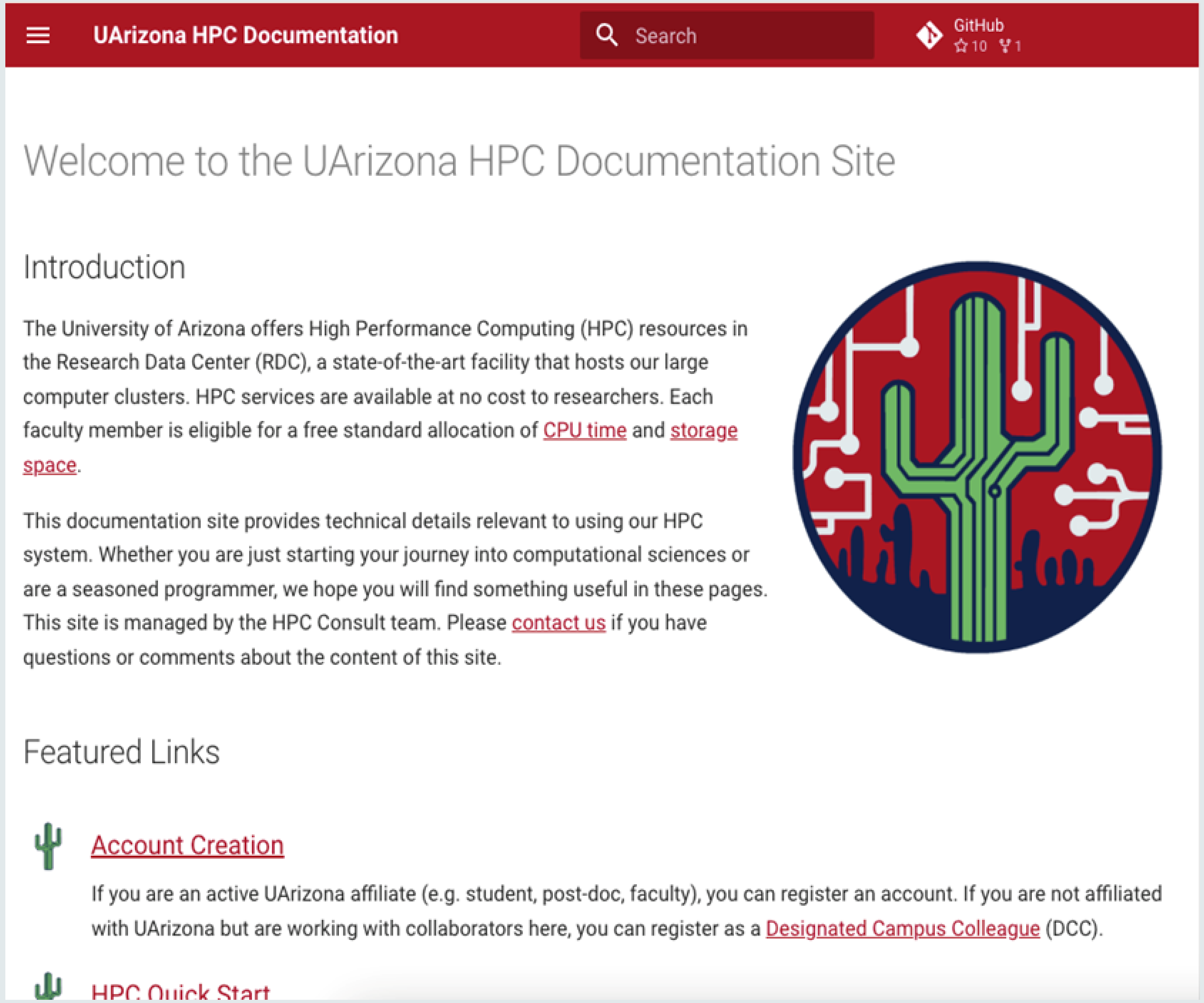
Remote computing

High Performance Computer (HPC)

Storage

There are a number of ways one can approach storage on the HPC:

- Your own folder (in /home/): 50GB limit
- Your group (in /groups/): 500GB limit
- Your PI research (in /xdisk/): 20TB



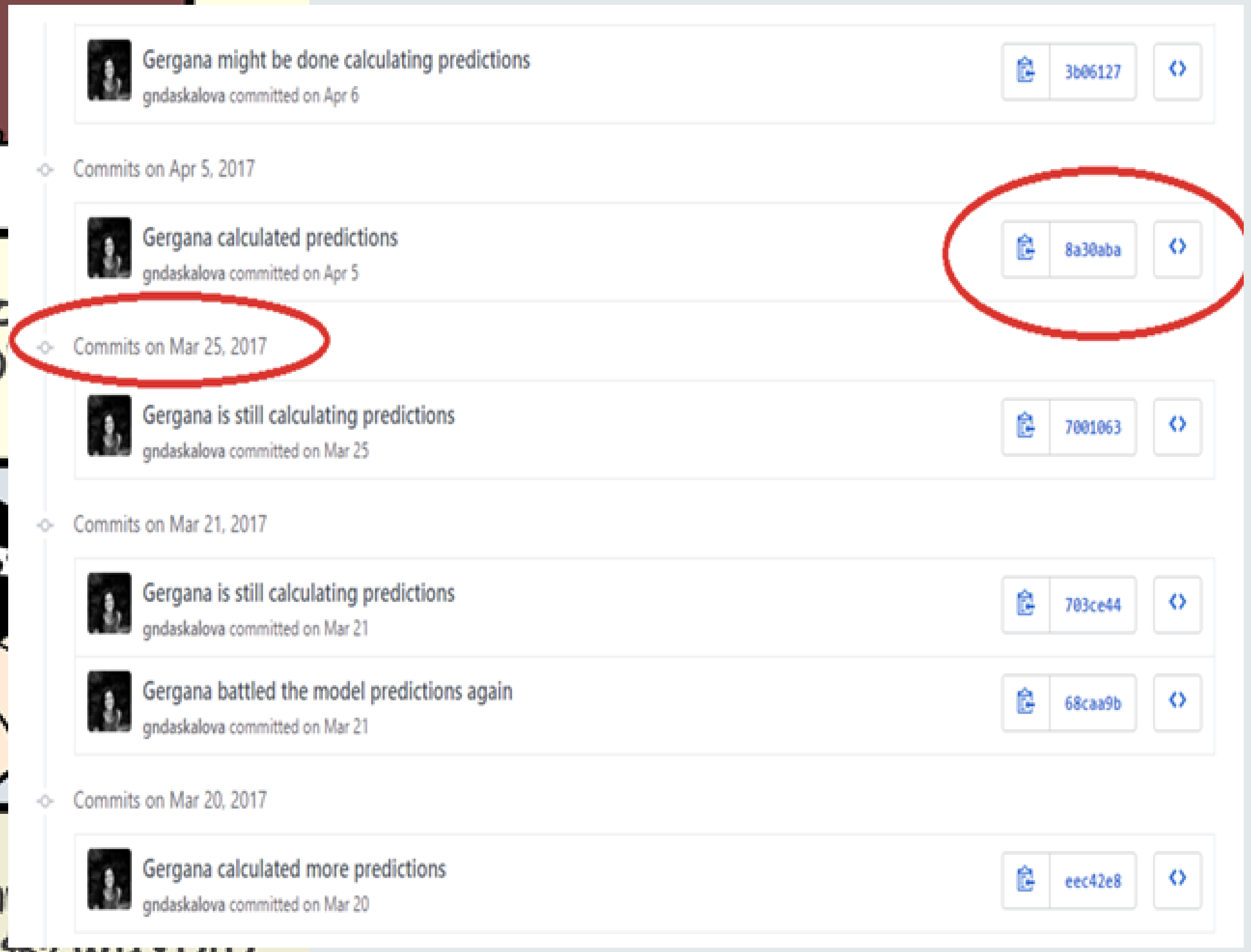
https://hpcdocs.hpc.arizona.edu/

Version control refers to keeping track of the version of a file, set of files, or a whole project.

Some version control tools:

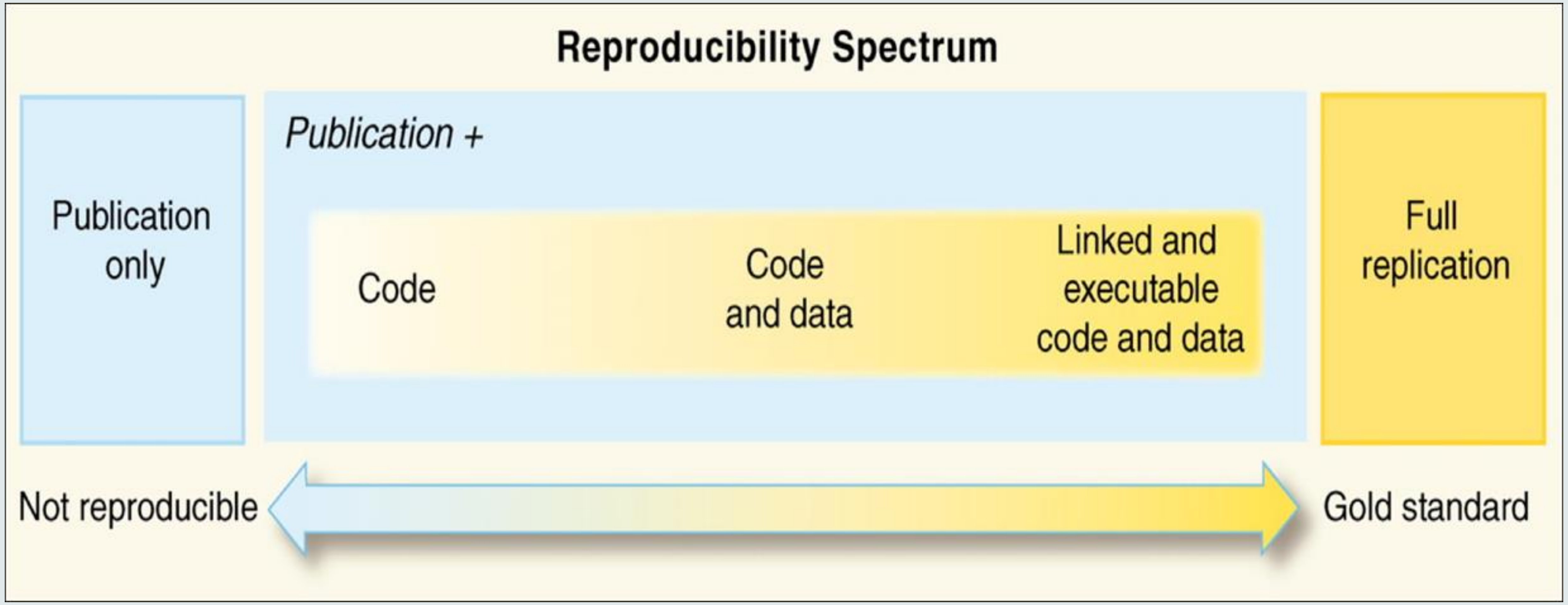
- Microsoft Office's *Track Changes* functionality
- Apple's *Time Machine*
- Google Docs' *Version History*
- Git

Example of the history for a repo with a R script inside it, as viewed on Github



Reproducibility

"Reproducing the result of a computation means running the same software on the same input data and obtaining the same results." Rougier et al. 2016



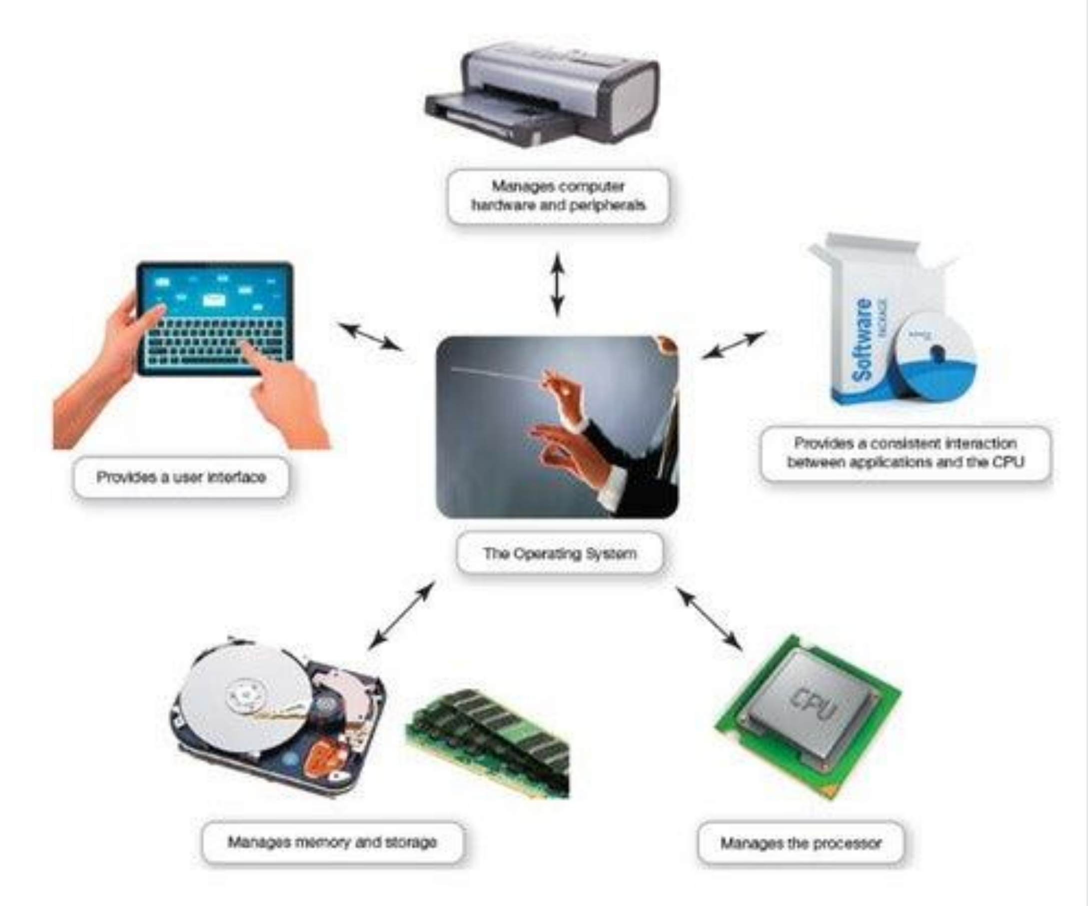
Reproducibility

1. Create a custom environment and share the recipe so your colleagues can replicate it on their computers
2. Package up the code and all the software and send it to your colleague as a Container.

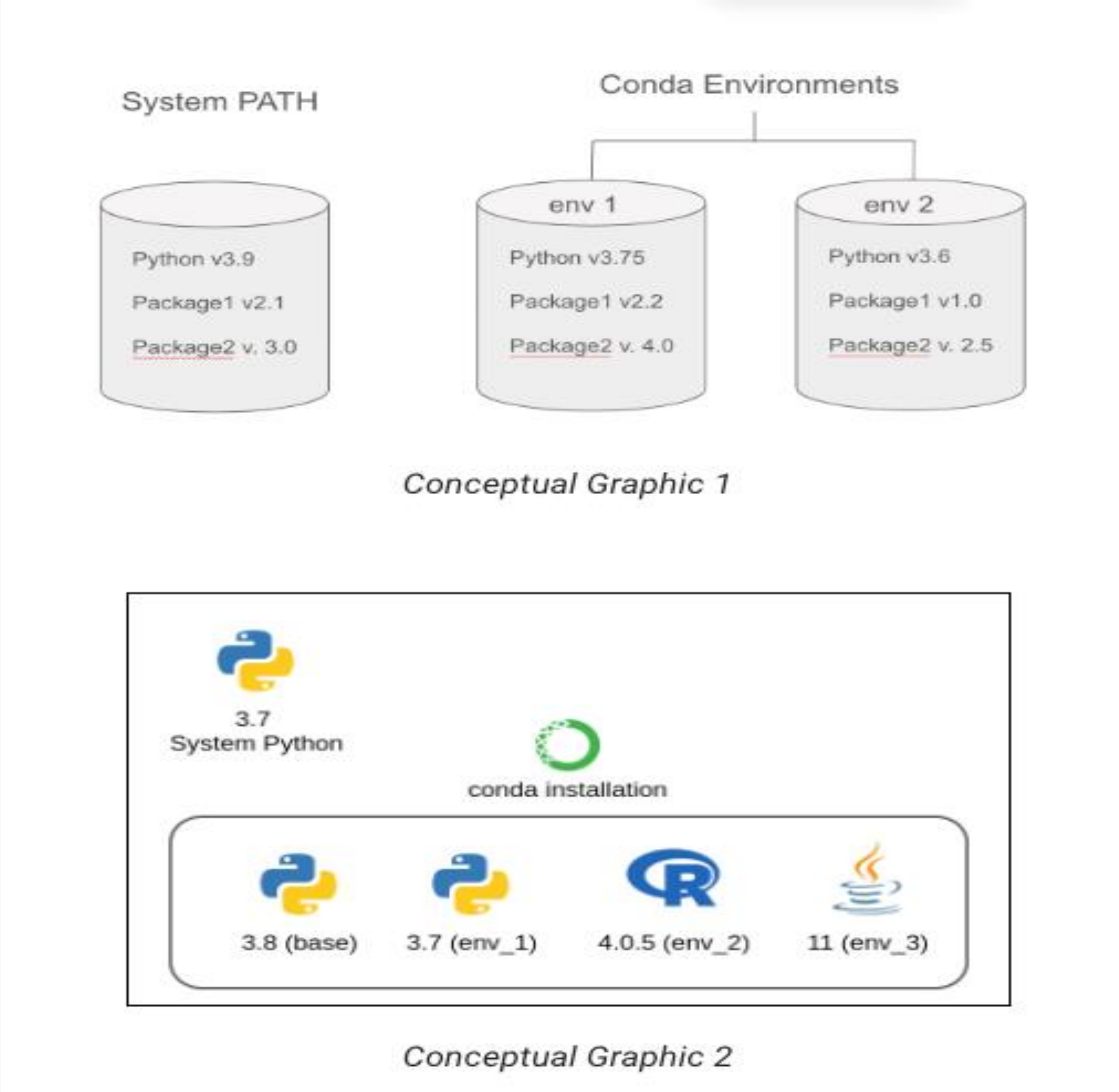
Reproducibility

A computing environment is the combination of hardware, software, and network resources that provide the infrastructure for computing operations and user interactions.

- **Hardware**: CPUs, GPUs, RAM
- **Operating system & version**: many flavors of Linux, MacOS, Windows
- **Software versions:** R, Python, etc.
- **Package versions:** specific R or Python packages, which often depend on other packages

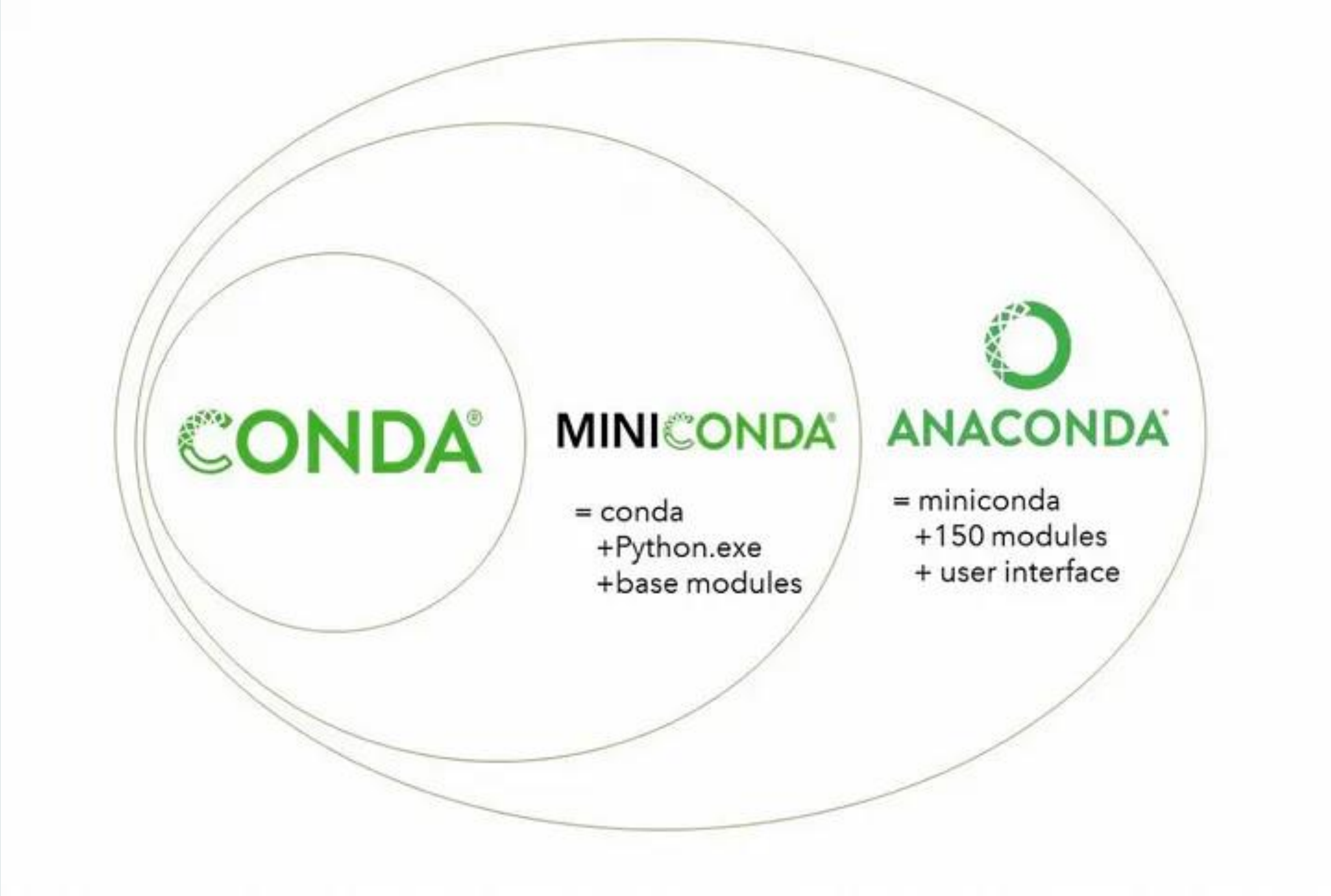


Reproducibility



Conceptual Graphic 1

Conceptual Graphic 2

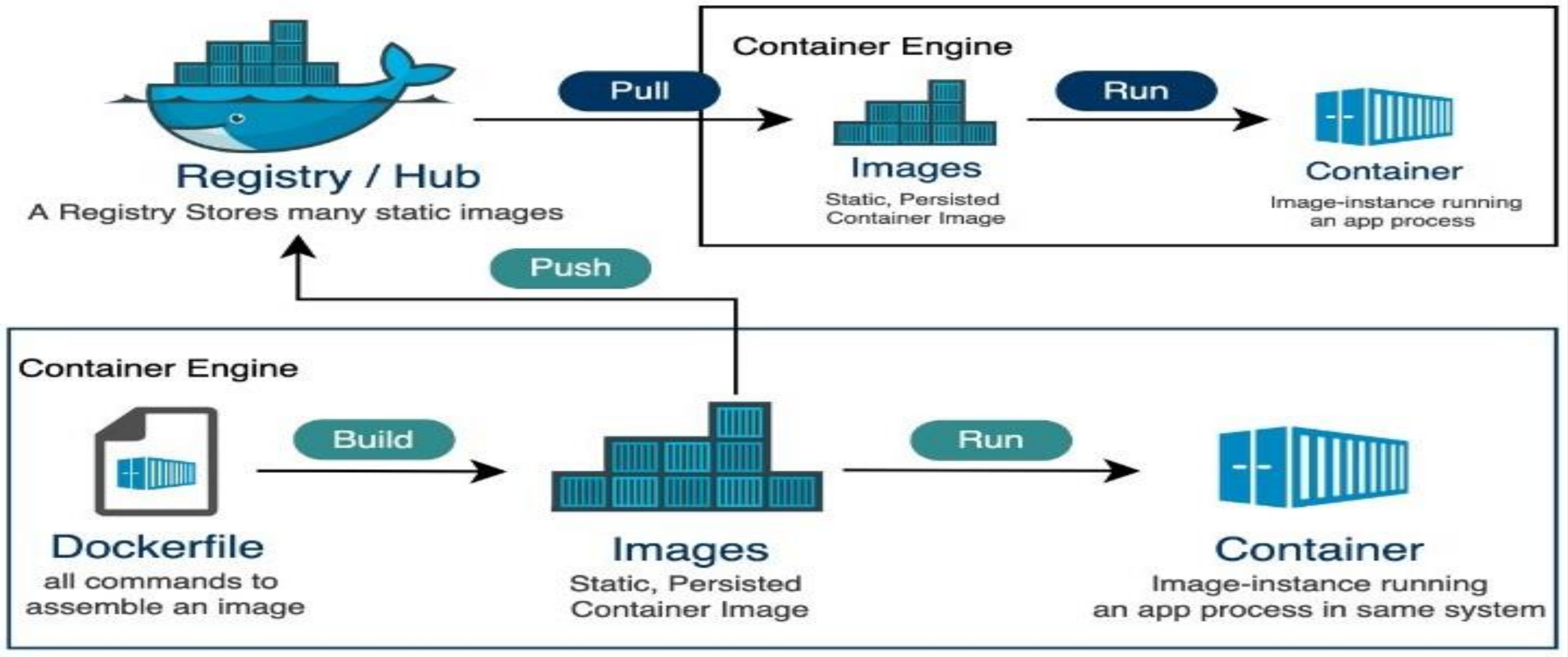


https://foss.cyverse.org/06_reproducibility_I/

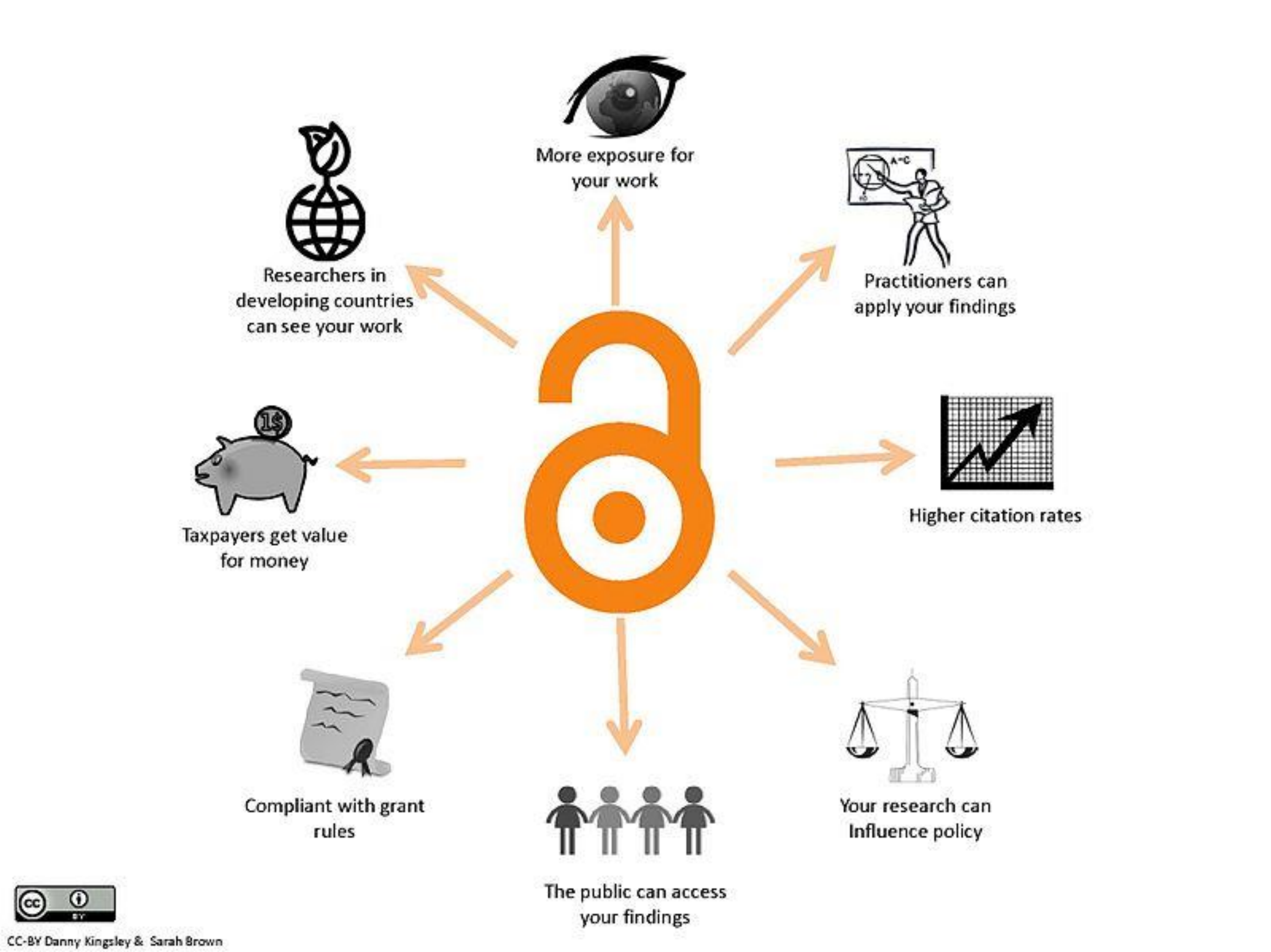
Reproducibility

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another.

Reproducibility



Benefits of Open Science

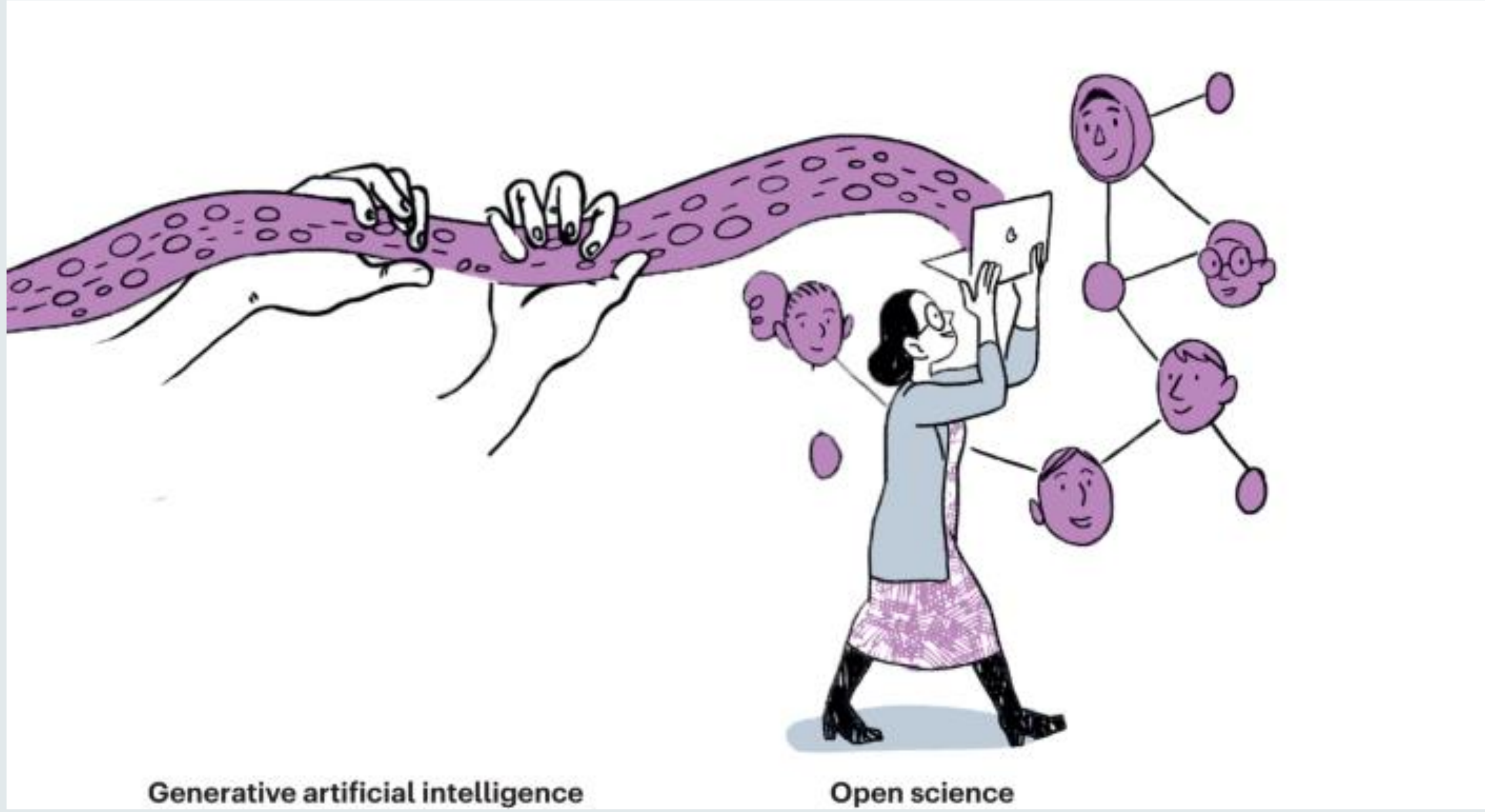


Challenges

- **Technological:** Low level of computing knowledge
- **Socio-cultural:** The lack of awareness of the benefits and importance of opening up research process
- **Organizational:** A closed culture is a challenge for individual researchers and slows down the overall openness of research

Challenges

- **Economic:** Resources and acceleration of innovations, significant investments
- **Legal:** Open science changes the way we look at ownership of data, copyright, privacy, and accountability in research.



Future Direction

- Scaling Open Science in Public Health
  - Vision for widespread adoption of FOSS principles.
  - Potential for new technologies like AI and machine learning in public health research.
- Training and Capacity Building
  - Importance of programs like FOSS for the next generation of public health researchers.

Conclusion

- Open science is a cornerstone for advancing public health by fostering transparency, reproducibility, and collaboration across disciplines.
- Open science can be taught in an open science curriculum / join the FOSS session
- Advocating for the adoption of open science practices in research communities

Useful Link

- FOSS Sessions: https://foss.cyverse.org/
- YouTube Channel: https://www.youtube.com/@CyverseOrgProject
- Cyverse Portal: https://user.cyverse.org/
- HPC: https://hpcdocs.hpc.arizona.edu/#introduction
- R4R: https://datascience.arizona.edu/r4r
- UofA DSI: https://datascience.arizona.edu/news

ccurley@arizona.edu

Contact Info

- Imran Mithu – Imranmithu@arizona.edu
- Caleigh Curley- ccurley@arizona.edu
- Joy Kinko Luzingu- joyluzingu@arizona.edu

Acknowledgment

RESEARCH, INNOVATION & IMPACT
Data Science Institute

CYVERSE®

THE UNIVERSITY OF ARIZONA
Mel & Enid Zuckerman
College of Public Health

Community, Environment & Policy

Epidemiology and Biostatistics


ANY
QUESTIONS?