



THE UNIVERSITY
OF ARIZONA

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)

What is R4R

Organizer



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

Roots for Resilience (R4R)



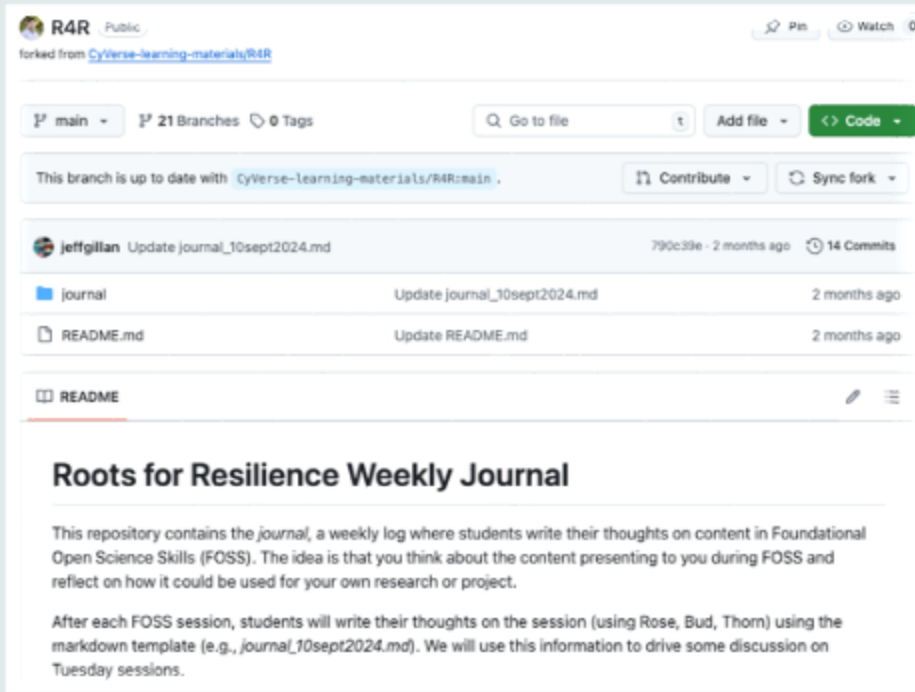
The Roots for Resilience Program provides training and support to select graduate students on open, reproducible science and computational infrastructure tools to enhance research focused on resiliency in the environment.

<https://datascience.arizona.edu/r4r>

What do we do during the fellowship?



Weekly Online FOSS Workshop on an Open Science Topic



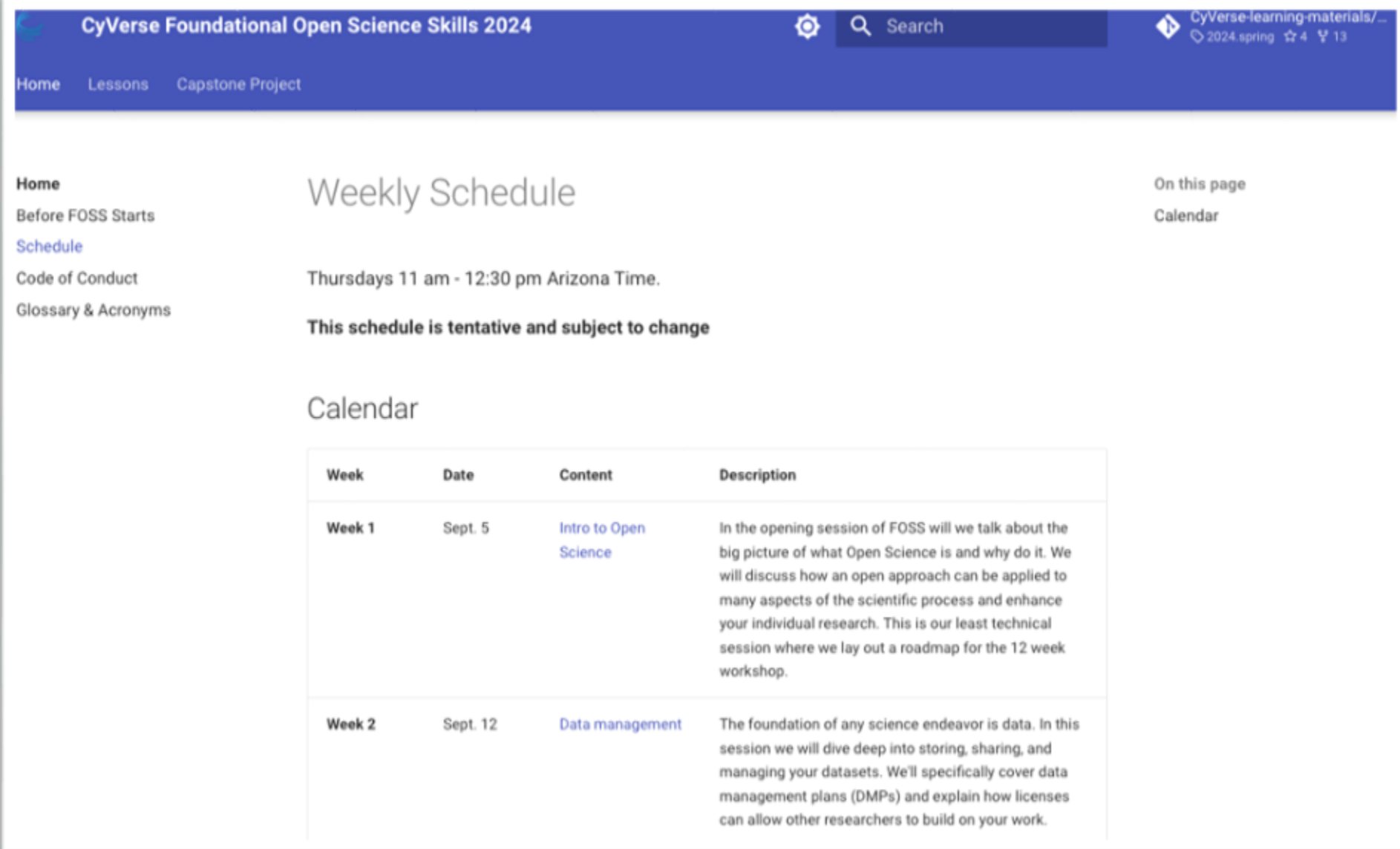
Weekly R4R Journal Submission



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)



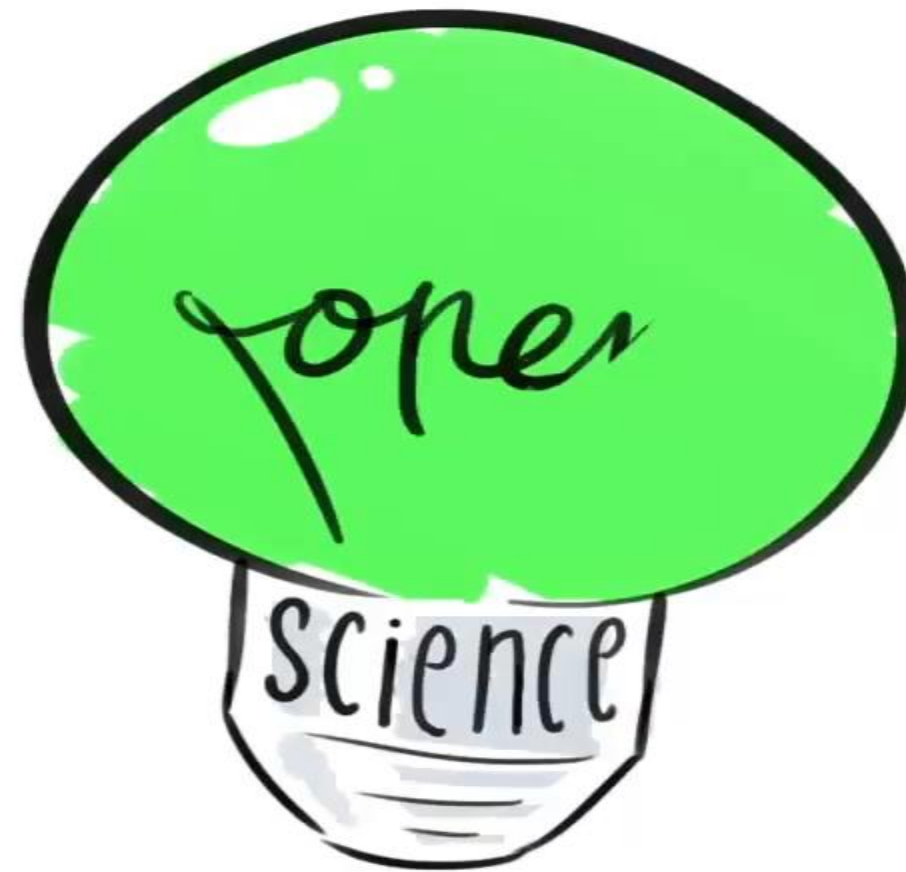
The screenshot shows the website for "CyVerse Foundational Open Science Skills 2024". The page title is "Weekly Schedule" and it indicates that sessions are held on Thursdays from 11 am to 12:30 pm Arizona Time. A note states that the schedule is tentative and subject to change. A calendar table lists the first two weeks of the training.

Week	Date	Content	Description
Week 1	Sept. 5	Intro to Open Science	In the opening session of FOSS we will talk about the big picture of what Open Science is and why do it. We will discuss how an open approach can be applied to many aspects of the scientific process and enhance your individual research. This is our least technical session where we lay out a roadmap for the 12 week workshop.
Week 2	Sept. 12	Data management	The foundation of any science endeavor is data. In this session we will dive deep into storing, sharing, and managing your datasets. We'll specifically cover data management plans (DMPs) and explain how licenses can allow other researchers to build on your work.

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

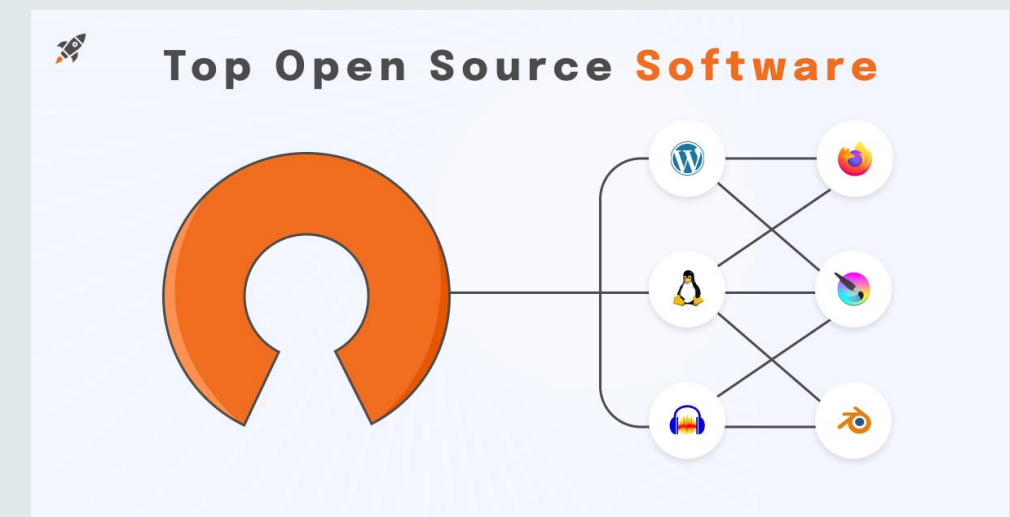
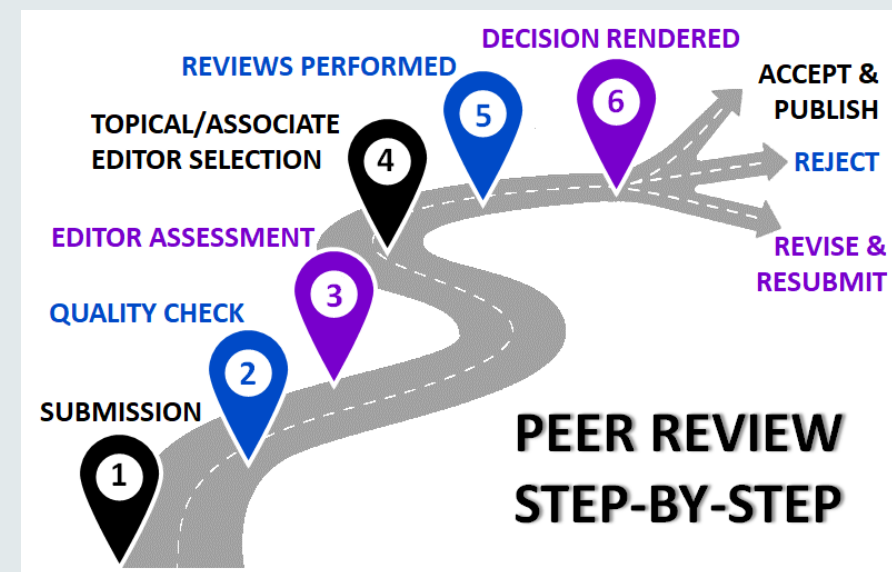
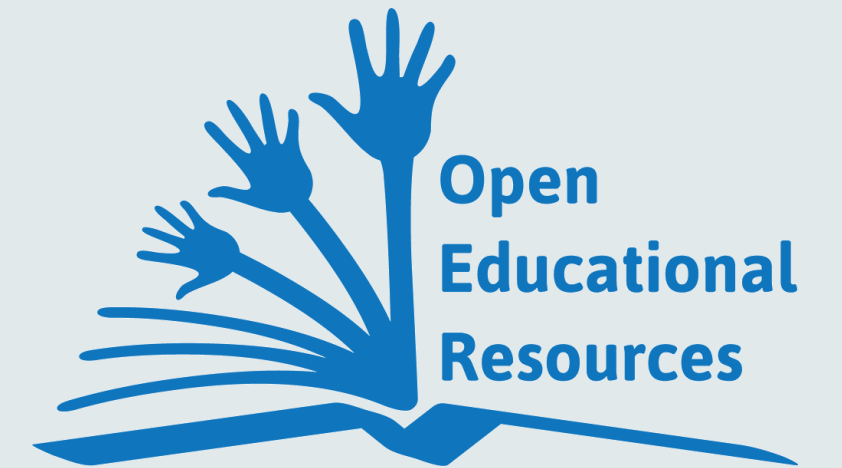
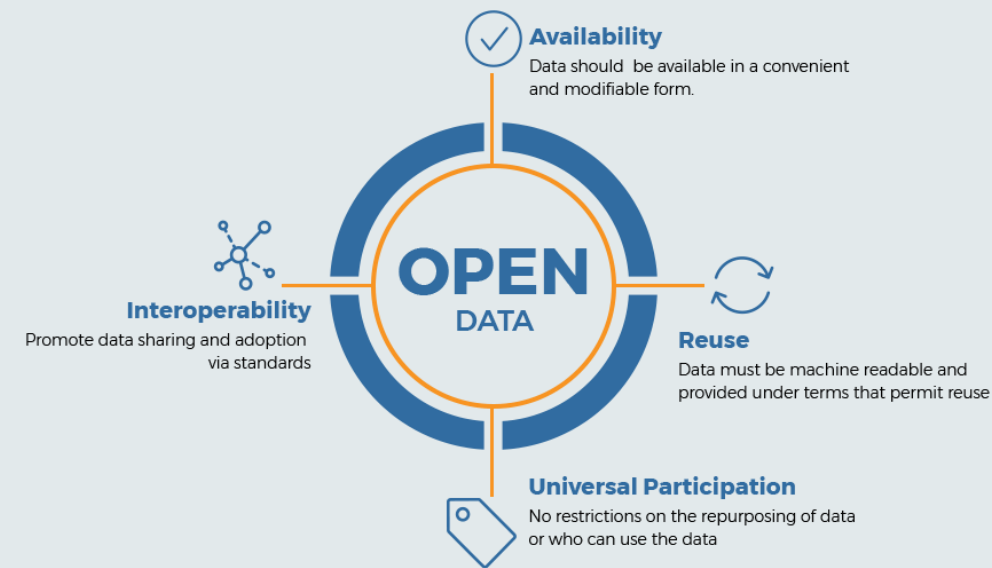
Overview of Open Science

What is Open Science



WHAT, HOW & WHY ?

Pillars of Open Science



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

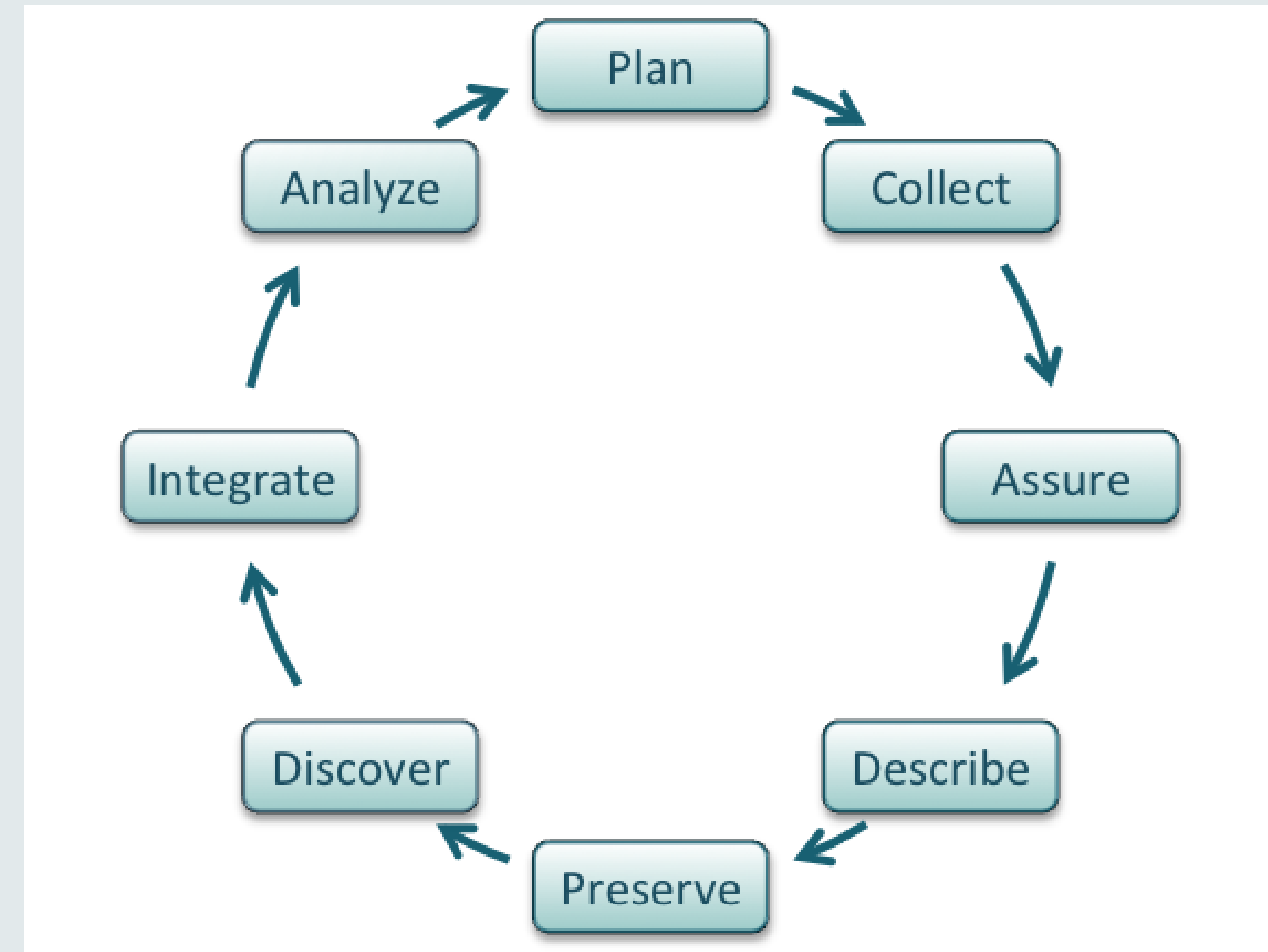
Open Science Framework



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

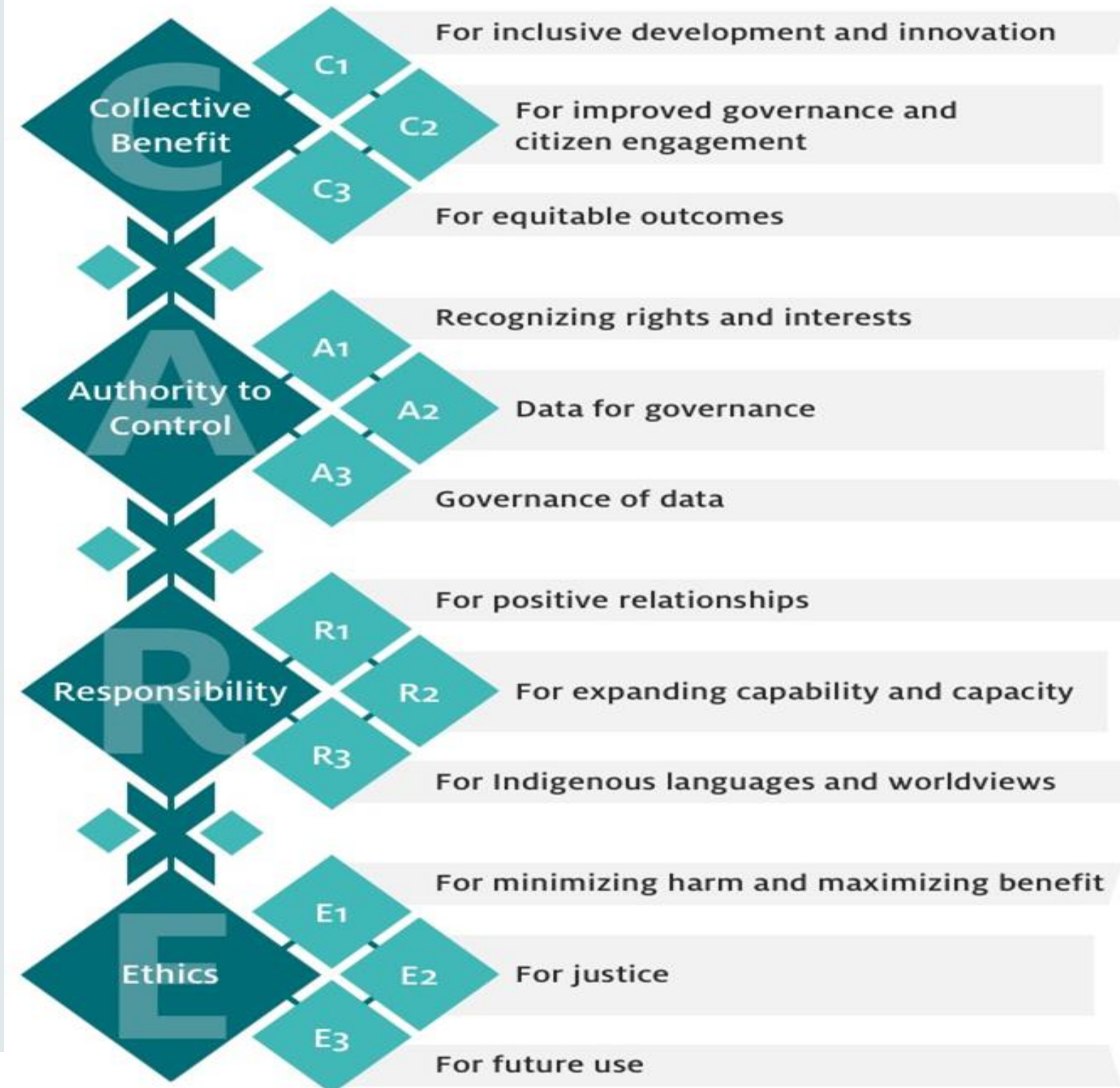
INDIGENOUS				MAINSTREAM		
New Zealand Indigenous Data Sovereignty Principles	Australia Indigenous Data Sovereignty Protocols	United States Indigenous Data Governance Principles	Canada Indigenous Data Governance Principles	Open Data Charter Principles	FAIR Principles for Data Management and Stewardship	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

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

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

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

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



Be

F A I R

Findable Accessible Interoperable Reusable

and

C A R E

**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

caleighcurley.github.io/caleigh-curley/

Caleigh's Homepage Publications Teaching Talks CV Blog Posts



Caleigh Curley
she/her/hers
DrPH Doctoral Candidate |
Research Program Administrator


Phoenix, AZ, USA
University of Arizona
Website
Email
Google Scholar
Semantic Scholar
ORCID
PubMed
ResearchGate
Github

Hello/Yá'át'ééh/Lolama!

My name is Caleigh Curley. I am a citizen of the Diné and Hopi Tribe; raised in Ganado, Arizona, within the Navajo Nation and now living on O'odam lands as a Doctoral Candidate in the [Public Health Policy and Management Program](#) at the University of Arizona (UArizona) Mel and Enid Zuckerman College of Public Health (MEZCOPH) and a member of the [Collaboratory for Indigenous Data Governance Lab](#). I earned a Bachelor of Science in Biomedical Sciences and a Master of Public Health in Health Promotion from Northern Arizona University. My research explores the intersection of Indigenous health, public health policy, and data governance. I have over six years of experience working with Tribes, Tribal colleges and universities, Tribal organizations, and urban Indian organizations. I provide expert analysis, advice, and guidance on health policy and analysis in my DrPH program. I provide expert analysis, advice, and guidance on regulatory, and legislative issues that affect Indigenous Peoples.

joyluzingu.github.io/teaching/

Joy Luzingu About Me Publications Talks Teaching CV



Joy Luzingu

Teaching


[Graduate Teaching Assistant: ECOL 182 L](#)
University of Arizona, College of Science, 2023
Taught lectures and administered labs for ECOL 182 L: Introduction to Biology/ Laboratory

[Graduate Teaching Assistant: MCB 181 L](#)
University of Arizona, College of Science, 2022
Instructed and administered labs for ECOL 182 L: Introduction to Biology/ Laboratory

[Teaching Assistant: ECOL 182 L](#)
University of Arizona, College of Science, 2022
Instructed and administered labs for ECOL 182 L: Introduction to Biology/ Laboratory

imranmithu.github.io/publications/

Home CV Research Publications Teaching Talks Blog



Imran Hossain Mithu
Doctoral Student in
Environmental Health Science
at the Mel and Enid Zuckerman
College of Public Health,
University of Arizona

Tucson, AZ, USA
ResearchGate
LinkedIn
Github
Google Scholar
ORCID

Publications

[Hidden danger: The long-term effect of ultrafine particles on mortality and its sociodemographic disparities in New York State](#)

Published in *Journal of Hazardous Materials*, 2024

Recommended citation: Qi, Q., Yu, F., Nair, A. A., Lau, S. S., Luo, G., Mithu, I., ... & Lin, S. (2024). Hidden danger: The long-term effect of ultrafine particles on mortality and its sociodemographic disparities in New York State. *Journal of Hazardous Materials*, 471, 134317.

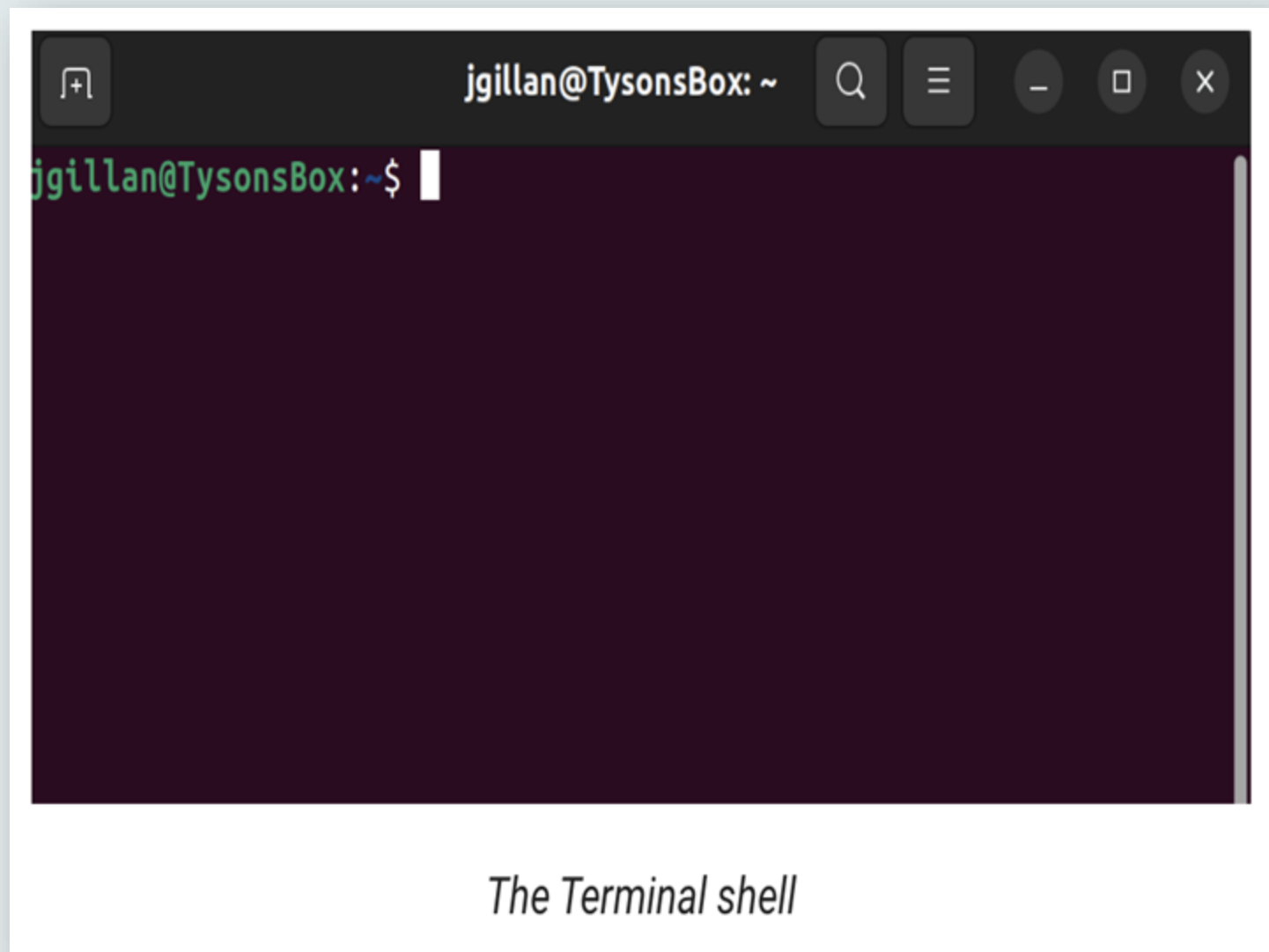
[The portrayal of antimicrobial resistance in Bangladeshi newspapers during 2010–2021: Toward understanding the narrative](#)

Published in *Plos one*, 2024

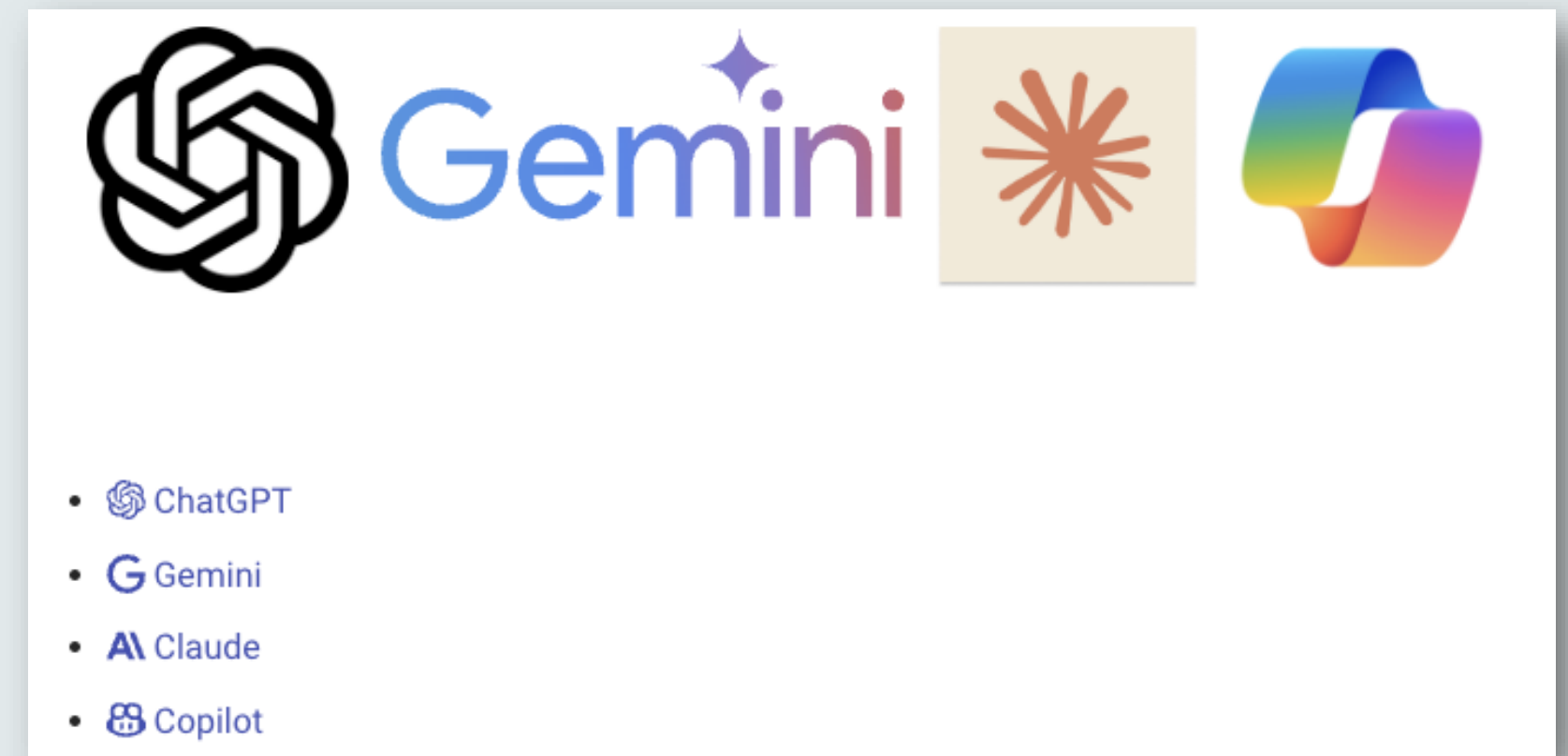
Recommended citation: Haque, T., Imtiaz, S. H., Hossain, M. I., Khan, S. H., Alam, M. M., Alam, Z., ... & Hassan, M. Z. (2024). The portrayal of antimicrobial resistance in Bangladeshi newspapers during 2010–2021: Toward understanding the narrative. *Plos one*, 19(5), e0304582.

How to talk to computers

The Command Line Interface (CLI)

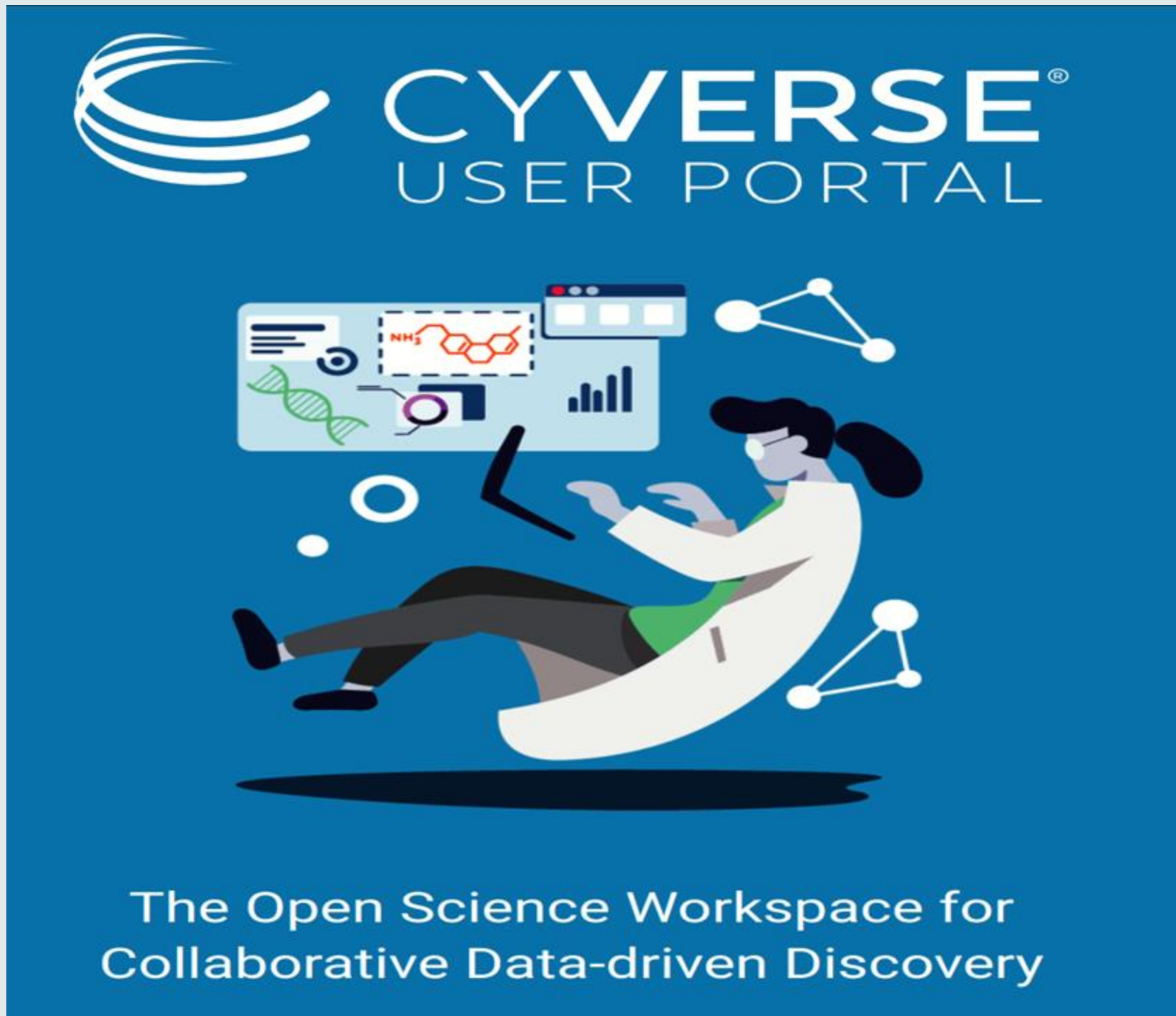


LLM Chatbots for Open Science



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

Remote computing



The image shows the CYVERSE USER PORTAL logo at the top left. Below it is an illustration of a scientist in a white lab coat sitting on the floor, working on a laptop. The scientist is surrounded by various data visualization icons, including a DNA double helix, a bar chart, a line graph, and a network diagram. The background is a solid blue color.

CYVERSE[®]
USER PORTAL

The Open Science Workspace for
Collaborative Data-driven Discovery

<https://cyverse.org/>

Remote computing

Resource Usage

Your current subscription tier is [Pro](#). [Buy](#)

Effective Period: 2024-09-03 19:13:13 - 2025-09-03 19:13:13

Data Storage
Using 0.00% as of 2024-09-17 14:28:04

CPU Consumption
Using 0.02% as of 2024-11-05 12:37:39

Analyses Stats
Total: 3

- Completed (3)
- Canceled (0)
- Failed (0)
- Submitted (0)
- Running (0)

Recent Analyses

- JupyterLab_Bioscience_analysis1**
Completed in 29 minutes - (2024-11-05 12:08:40)
No description was provided.
- Cloud_Shell_analysis1**
Completed in 1 minute - (2024-10-31 12:42:53)
No description was provided.
- Rocker_RStudio_Geospatial_analysis1**
Completed in 8 minutes - (2024-10-31 12:23:17)
No description was provided.

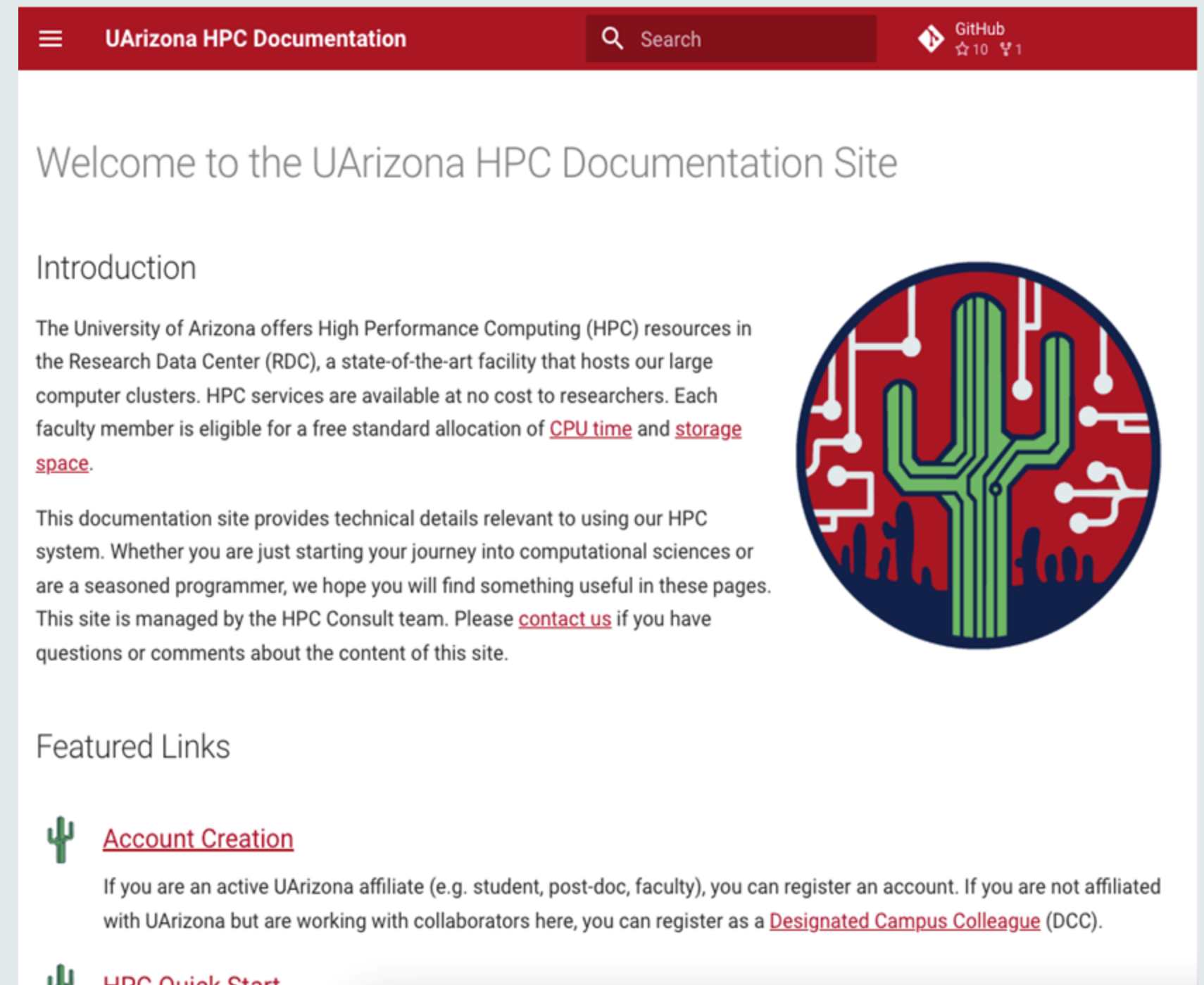
Remote computing

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



UArizona HPC Documentation



Welcome to the UArizona HPC Documentation Site

Introduction

The University of Arizona offers High Performance Computing (HPC) resources in the Research Data Center (RDC), a state-of-the-art facility that hosts our large computer clusters. HPC services are available at no cost to researchers. Each faculty member is eligible for a free standard allocation of [CPU time](#) and [storage space](#).

This documentation site provides technical details relevant to using our HPC system. Whether you are just starting your journey into computational sciences or are a seasoned programmer, we hope you will find something useful in these pages. This site is managed by the HPC Consult team. Please [contact us](#) if you have questions or comments about the content of this site.

Featured Links

-  [Account Creation](#)
If you are an active UArizona affiliate (e.g. student, post-doc, faculty), you can register an account. If you are not affiliated with UArizona but are working with collaborators here, you can register as a [Designated Campus Colleague \(DCC\)](#).
-  [HPC Quick Start](#)

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

"FINAL".doc



FINAL.doc!



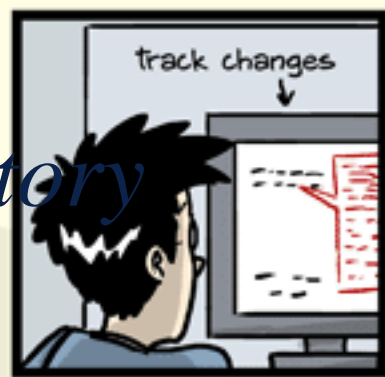
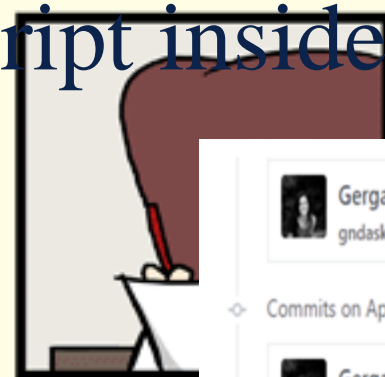
FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.c
CORRECTIO

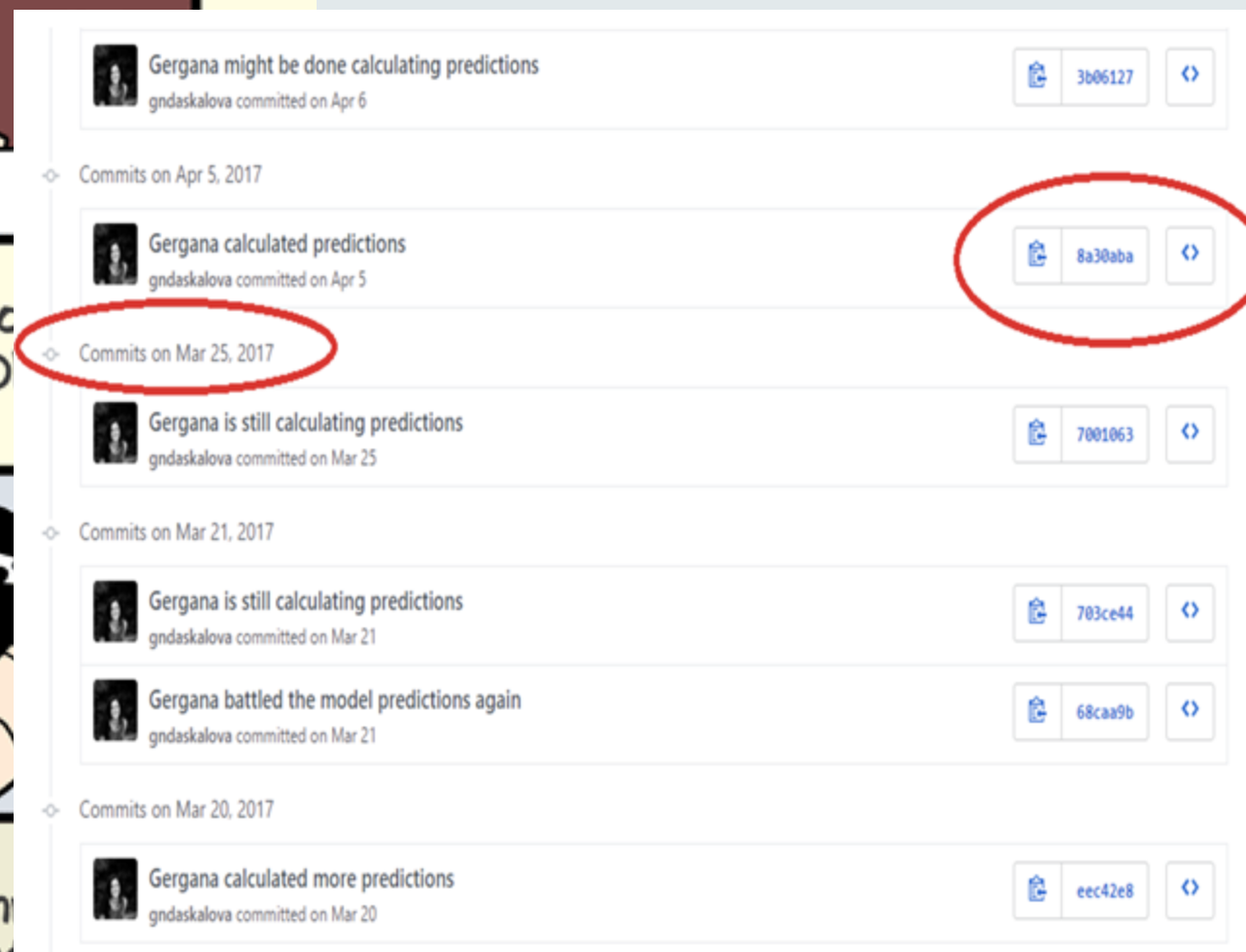


FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.com
corrections.10.#@%*WHYDID
ICOMETOGRADSCHOOL?????.doc

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



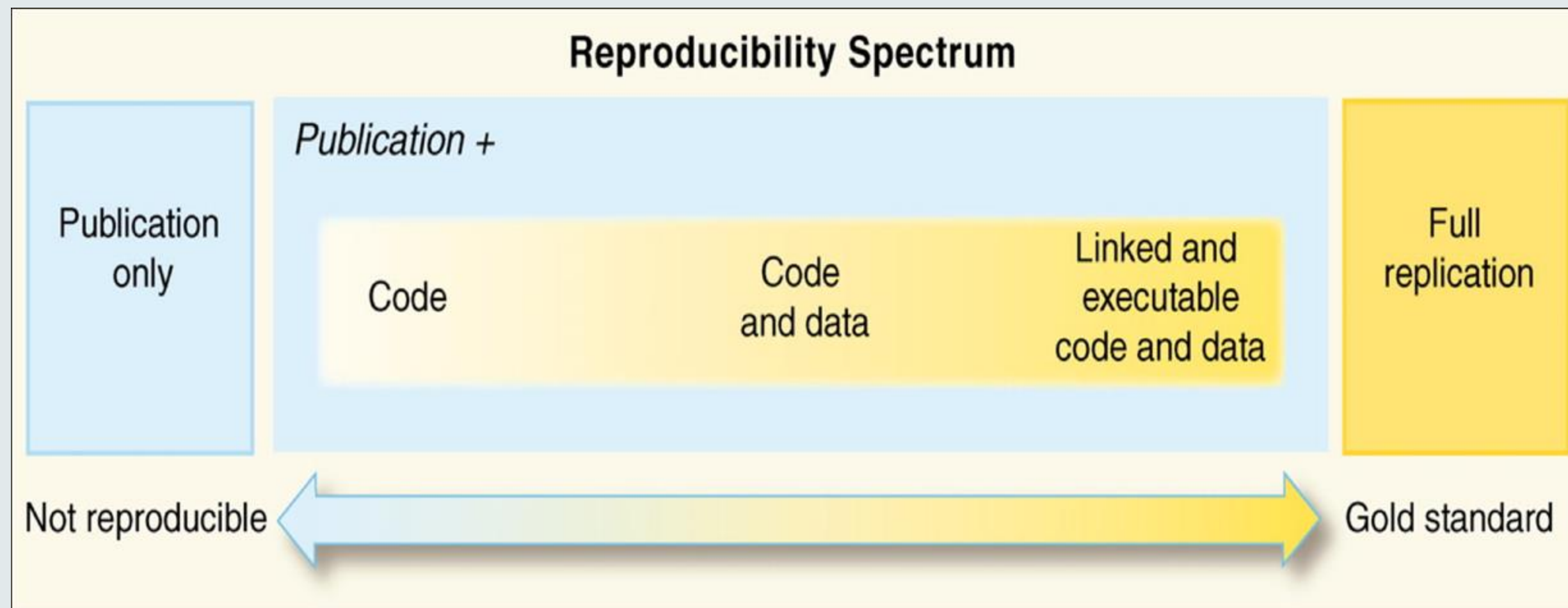
Version control refers to keeping track of the version of a file, set of files, or a whole project.

Some version control tools:

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

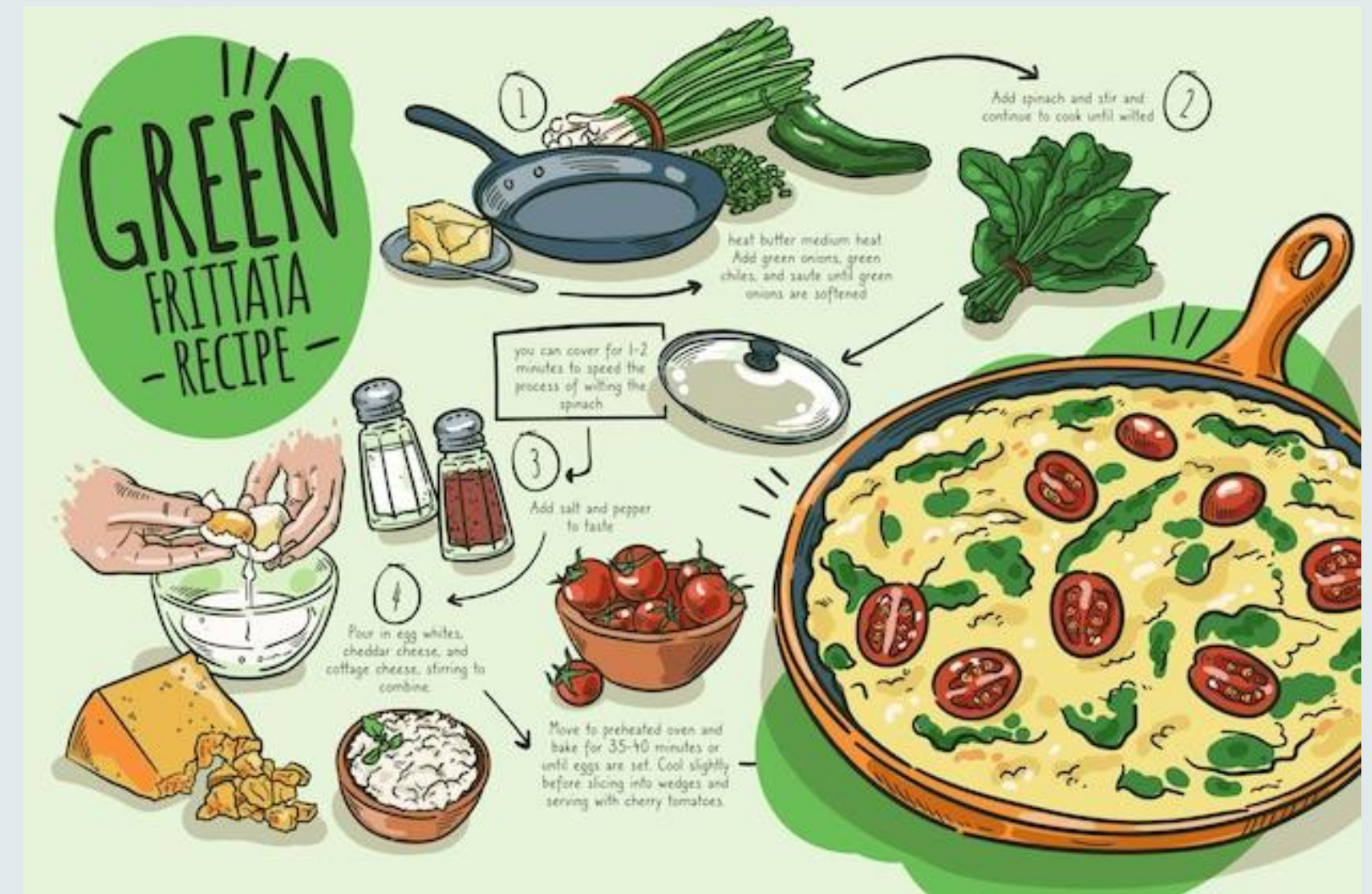
Reproducibility

"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

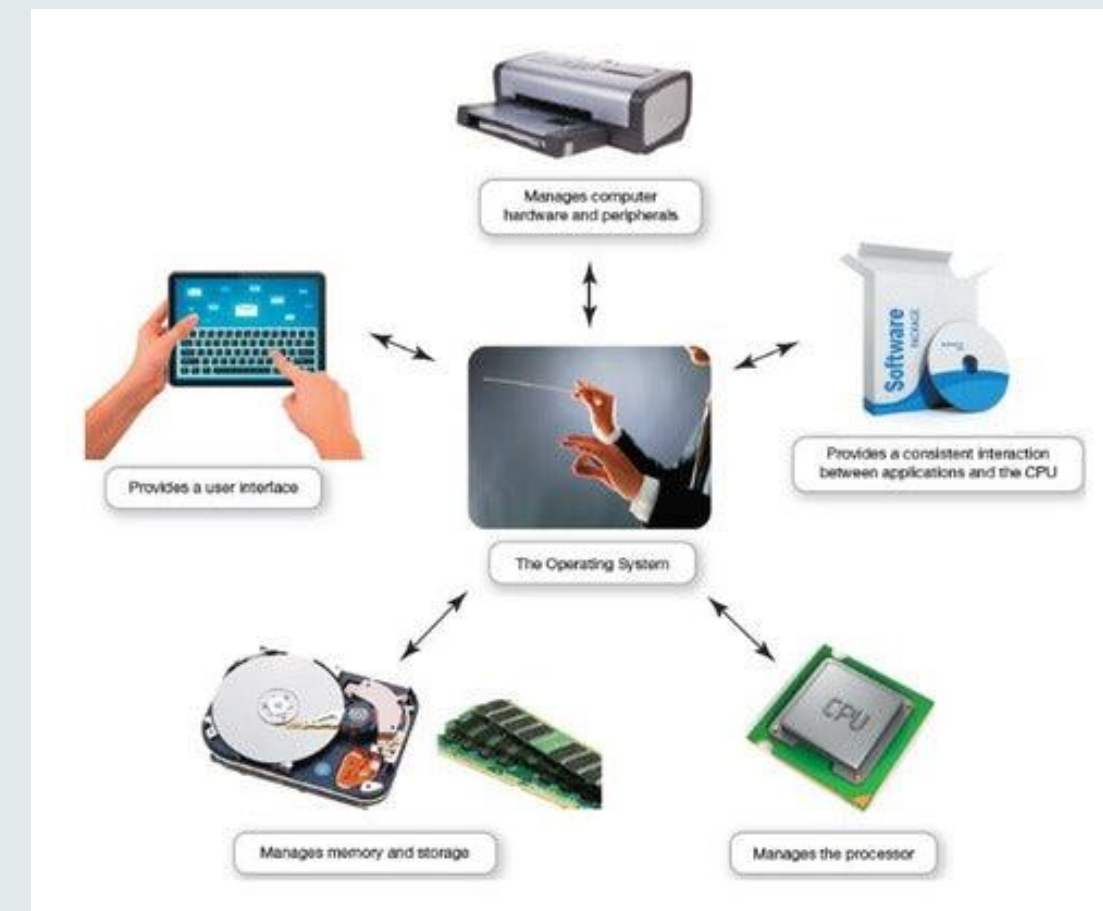
1. Create a custom environment and share the recipe so your colleagues can replicate it on their computers
2. Package up the code and all the software and send it to your colleague as a Container.



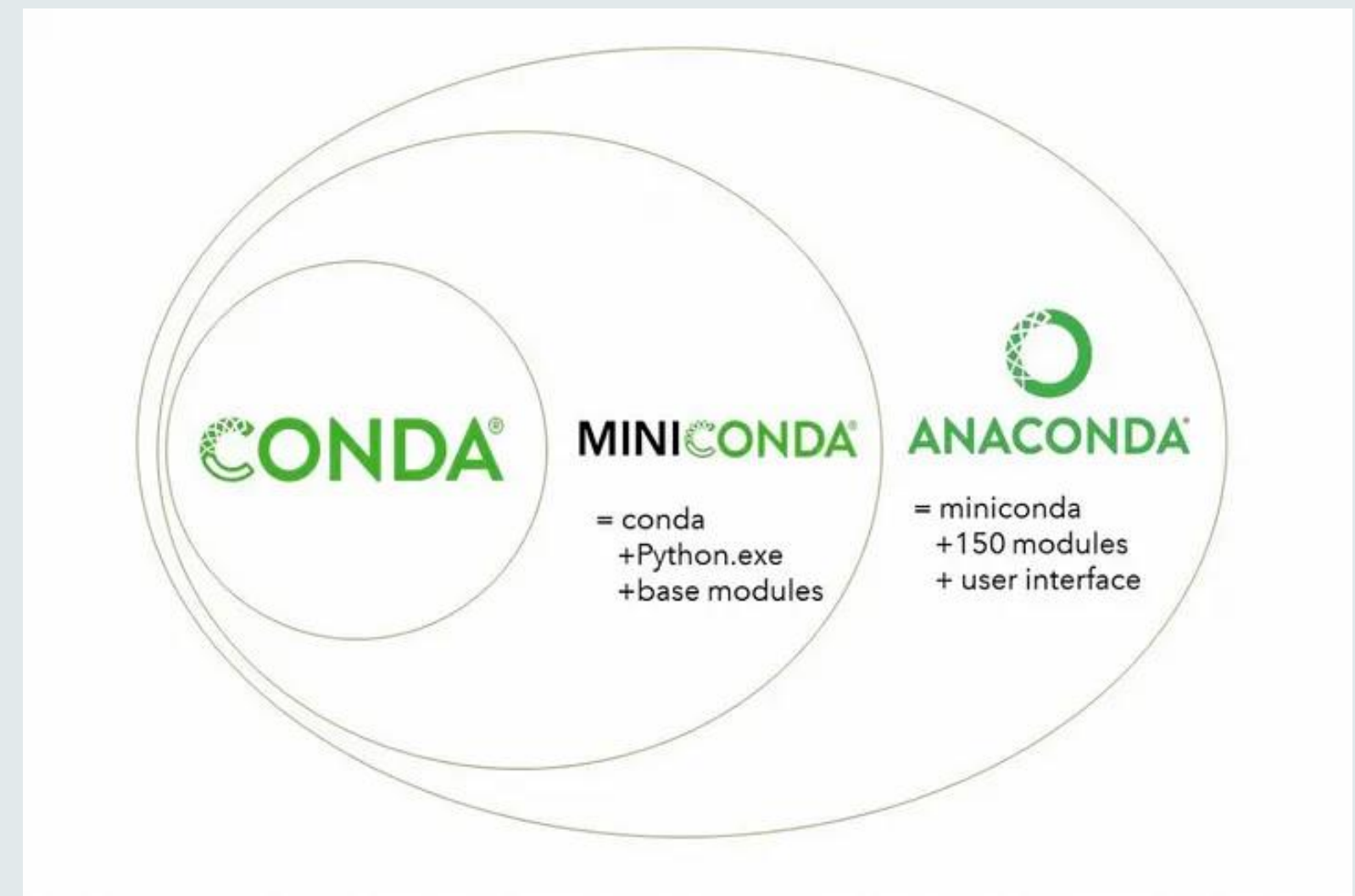
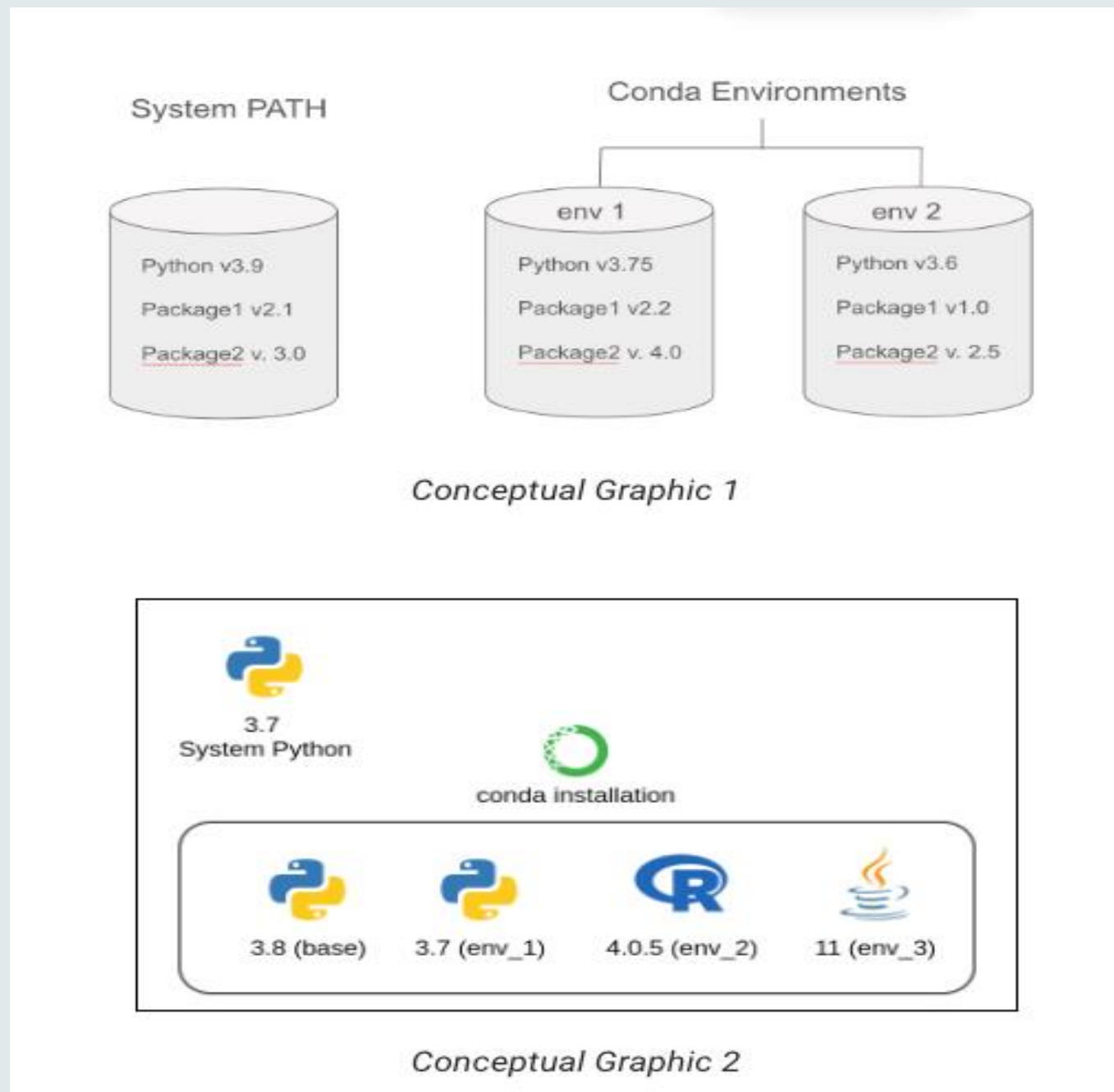
Reproducibility

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



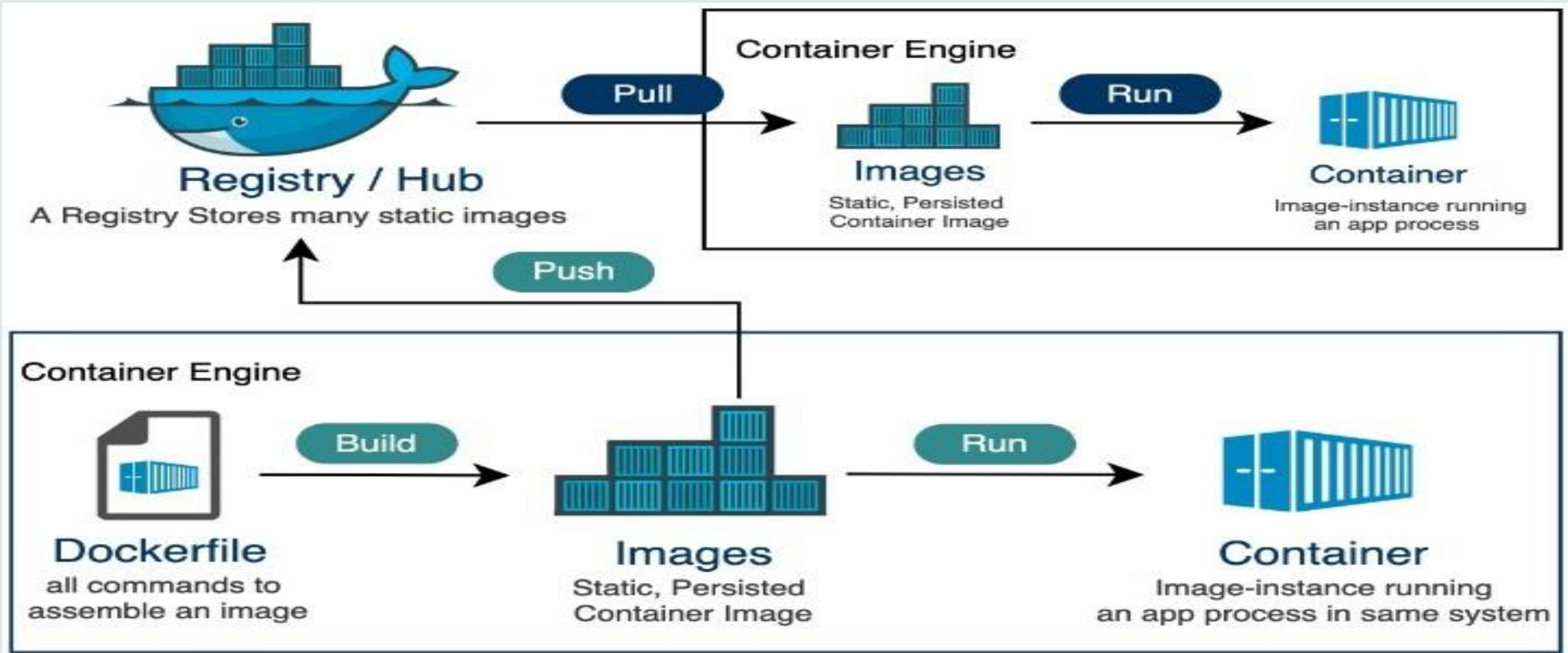
https://foss.cyverse.org/06_reproducibility_I/

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



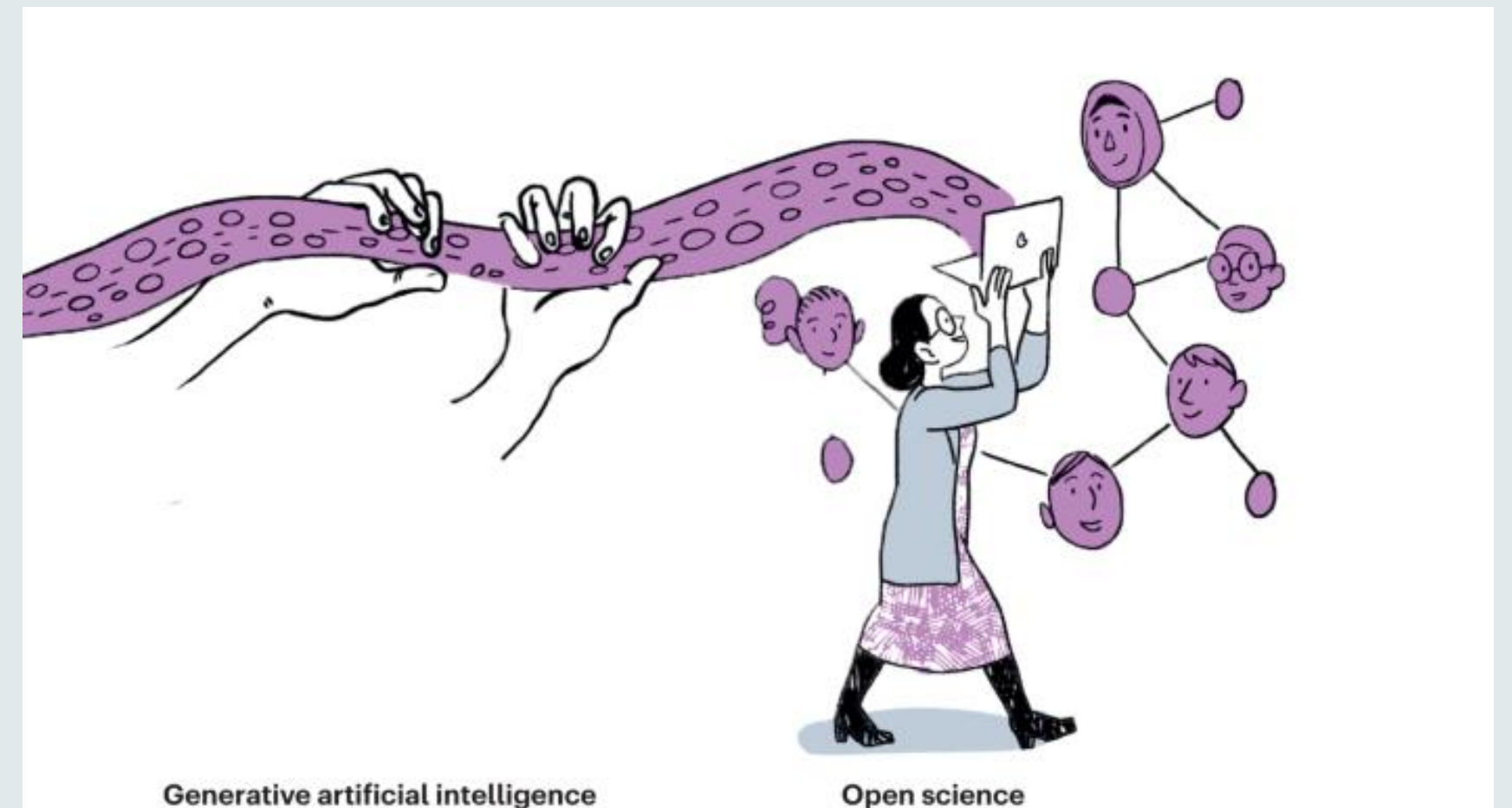
Challenges

- **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>

Contact Info

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

Acknowledgment



Community, Environment & Policy

Epidemiology and Biostatistics

ANY
QUESTIONS?

