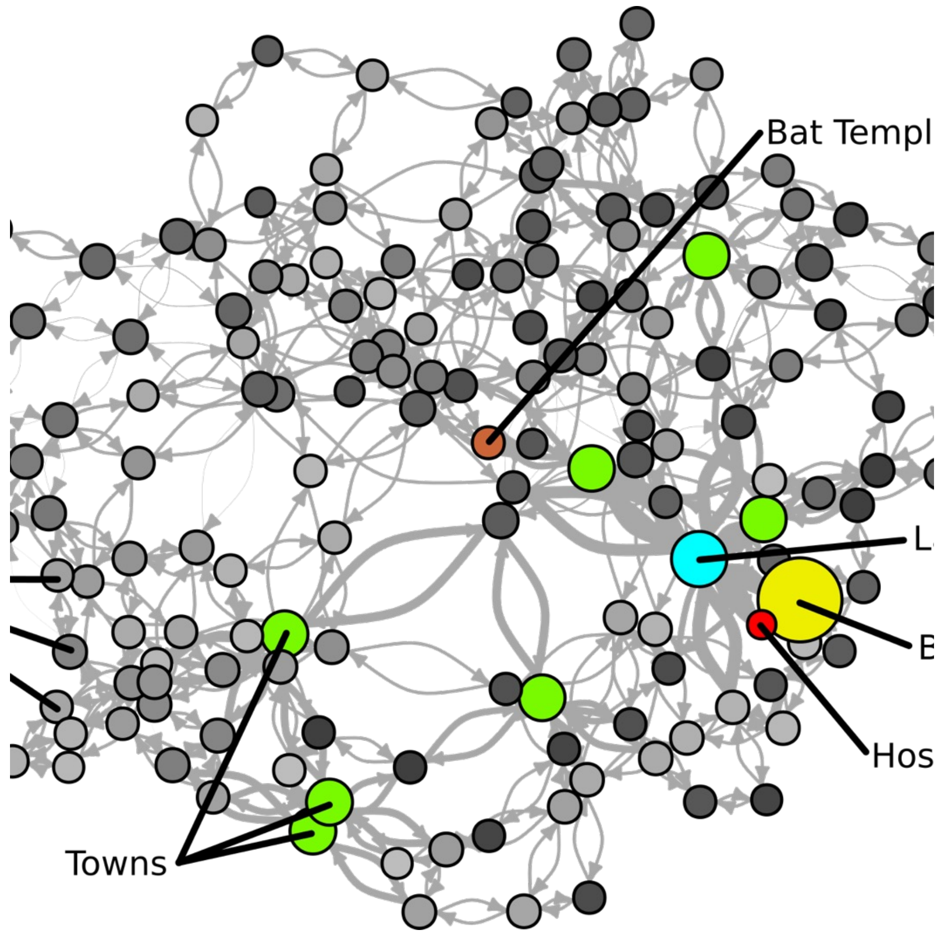
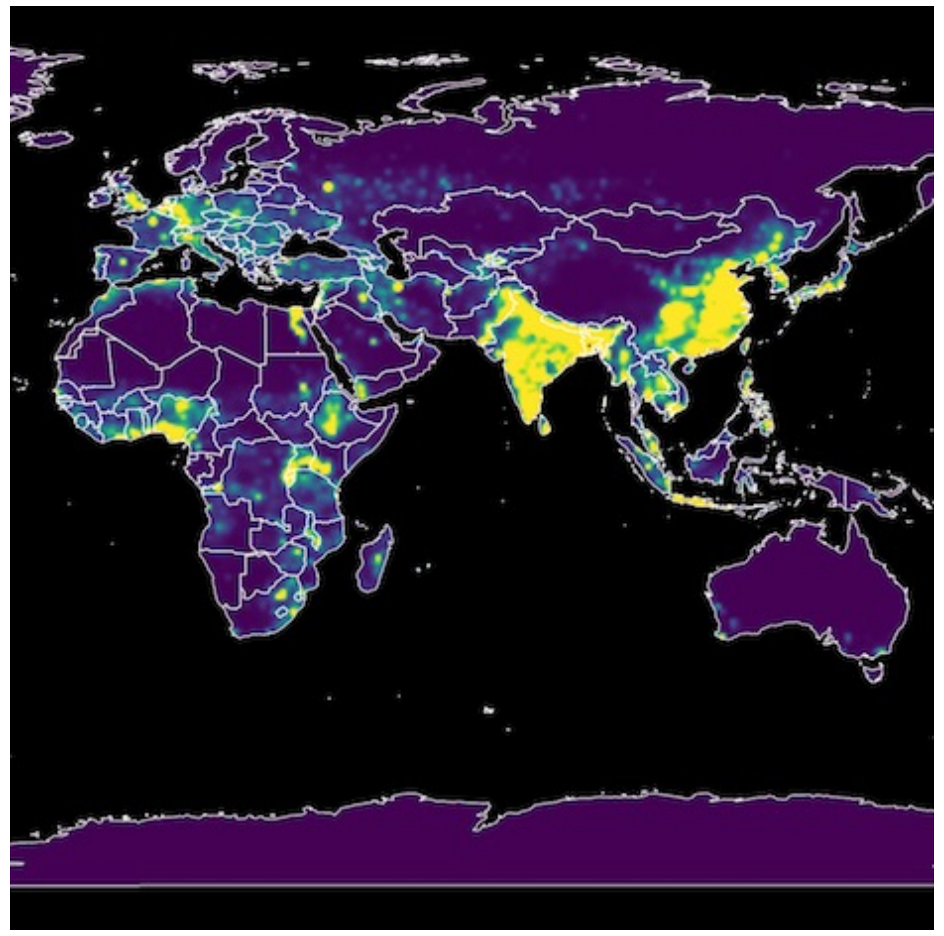
Tools and Teams for Reproducible Science with R

Noam Ross, EcoHealth Alliance

R User Group at the Harvard Data Science Initiative
2021-07-22





EcoHealth Alliance

@EcoHealthNYC ecohealthalliance.org

Building technical and community infrastructure for R to support open, reproducible science

@rOpenSci@hachyderm.io

ropensci.org

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

Article | **Authors** | **Metrics** | **Comments** | **Media Coverage**

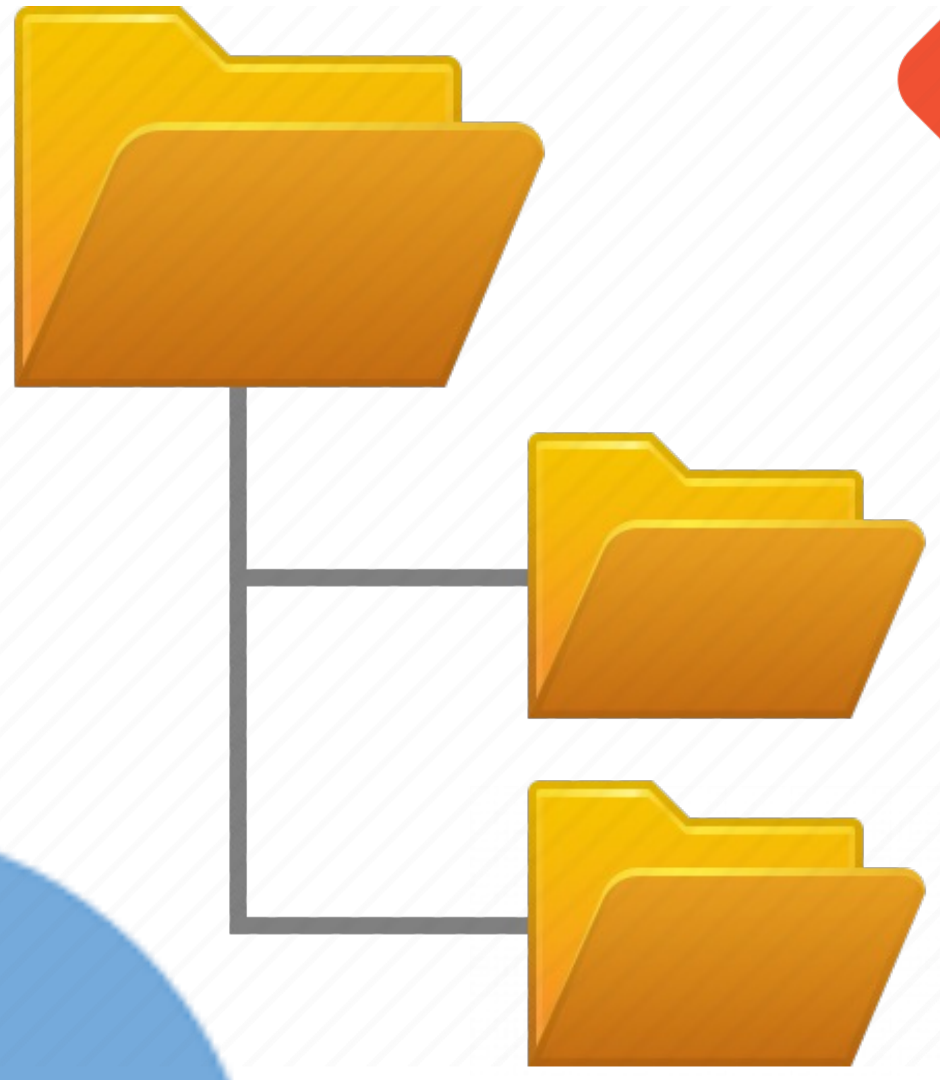
Author summary

Computers are now essential in all branches of science, but most researchers are never taught the equivalent of basic lab skills for research computing. As a result, data can get lost, analyses can take much longer than necessary, and researchers are limited in how effectively they can work with software and data. Computing workflows need to follow the same practices as lab projects and notebooks, with organized data, documented steps, and the project structured for reproducibility, but researchers new to computing often don't know where to start. This paper presents a set of good computing practices that every researcher can adopt, regardless of their current level of computational skill. These practices, which encompass data management, programming, collaborating with colleagues, organizing projects, tracking work, and writing manuscripts, are drawn from a wide variety of published sources from our daily lives and from our work with volunteer organizations that have delivered workshops to over 11,000 people since 2010.

Box 1. Summary of practices

1. Data management
   a. Save the raw data.
   b. Ensure that raw data are backed up in more than one location.
   c. Create the data you wish to see in the world.
   d. Create analysis-friendly data.
   e. Record all the steps used to process data.
   f. Anticipate the need to use multiple tables, and use a unique identifier for every record.
   g. Submit data to a reputable DOI-issuing repository so that others can access and cite it.
2. Software
   a. Place a brief explanatory comment at the start of every program.
   b. Decompose programs into functions.
   c. Be ruthless about eliminating duplication.
   d. Always search for well-maintained software libraries that do what you need.
   e. Test libraries before relying on them.
   f. Give functions and variables meaningful names.
   g. Make dependencies and requirements explicit.
   h. Do not comment and uncomment sections of code to control a program's behavior.
   i. Provide a simple example or test data set.
   j. Submit code to a reputable DOI-issuing repository.
3. Collaboration
   a. Create an overview of your project.
   b. Create a shared "to-do" list for the project.
   c. Decide on communication strategies.
   d. Make the license explicit.
   e. Make the project citable.
4. Project organization
   a. Put each project in its own directory, which is named after the project.
   b. Put text documents associated with the project in the `doc` directory.
   c. Put raw data and metadata in a data directory and files generated during cleanup and analysis in a results directory.
   d. Put project source code in the `src` directory.
   e. Put external scripts or compiled programs in the `bin` directory.
   f. Name all files to reflect their content or function.
5. Keeping track of changes
   a. Back up (almost) everything created by a human being as soon as it is created.
   b. Keep changes small.
   c. Share changes frequently.
   d. Create, maintain, and use a checklist for saving and sharing changes to the project.
   e. Store each project in a folder that is mirrored off the researcher's working machine.
   f. Add a file called `CHANGELOG.txt` to the project's `docs` subfolder.
   g. Copy the entire project whenever a significant change has been made.
   h. Use a version control system.
6. Manuscripts
   a. Write manuscripts using online tools with rich formatting, change tracking, and reference management.
   b. Write the manuscript in a plain text format that permits version control.

The "Good Enough" Toolkit



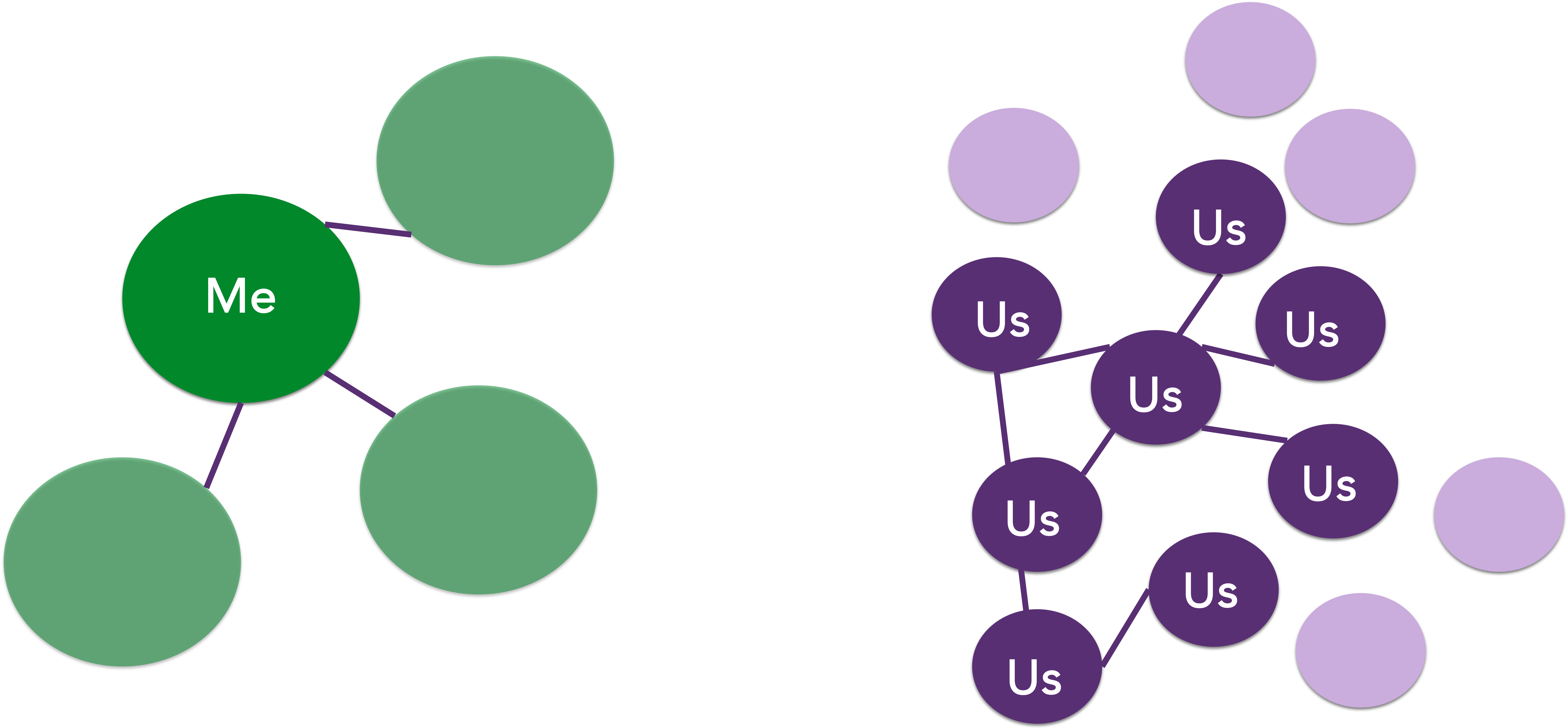
git

rmarkdown

GitHub

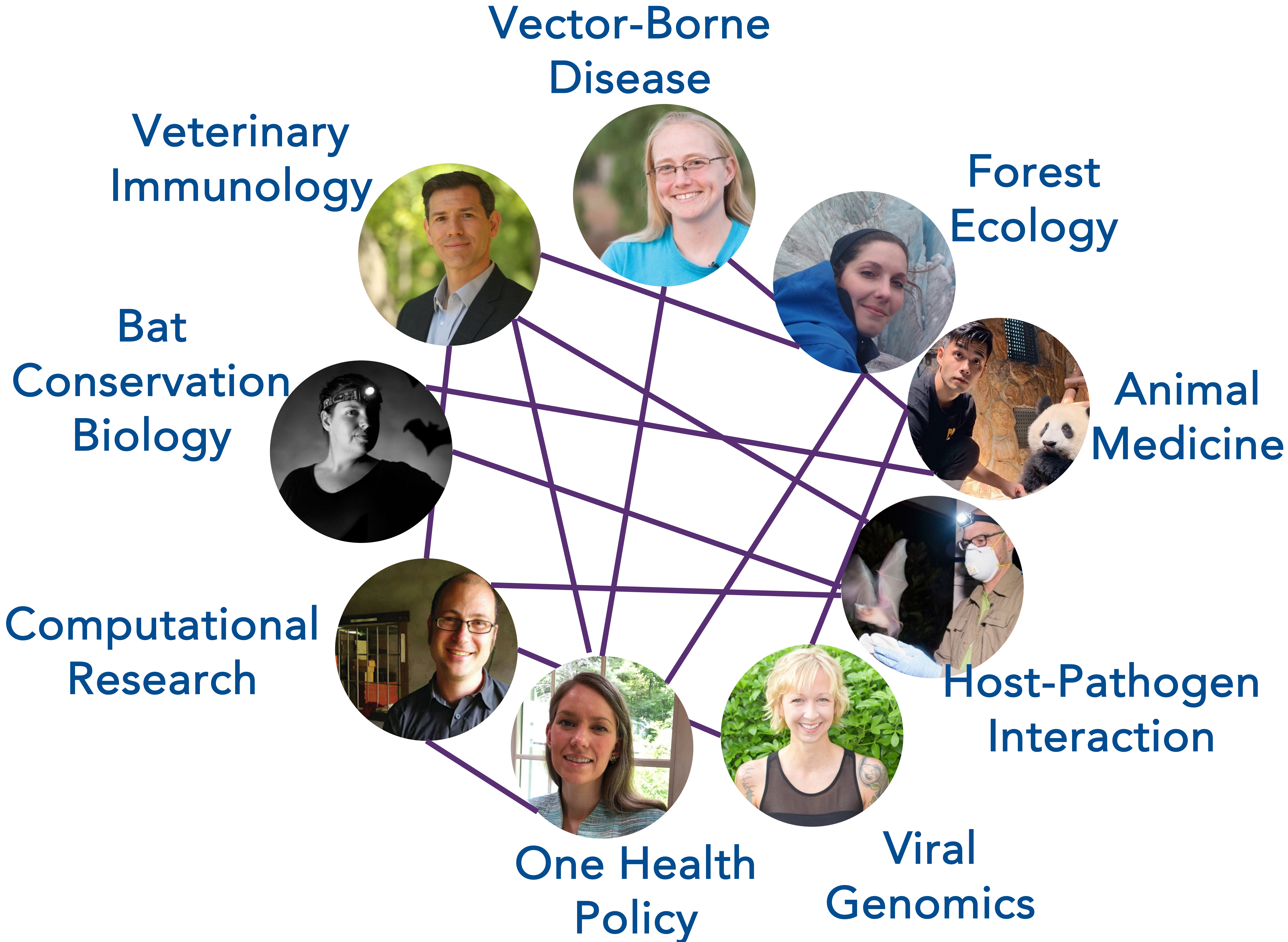
zenodo

How to Scale these Practices and Tools?



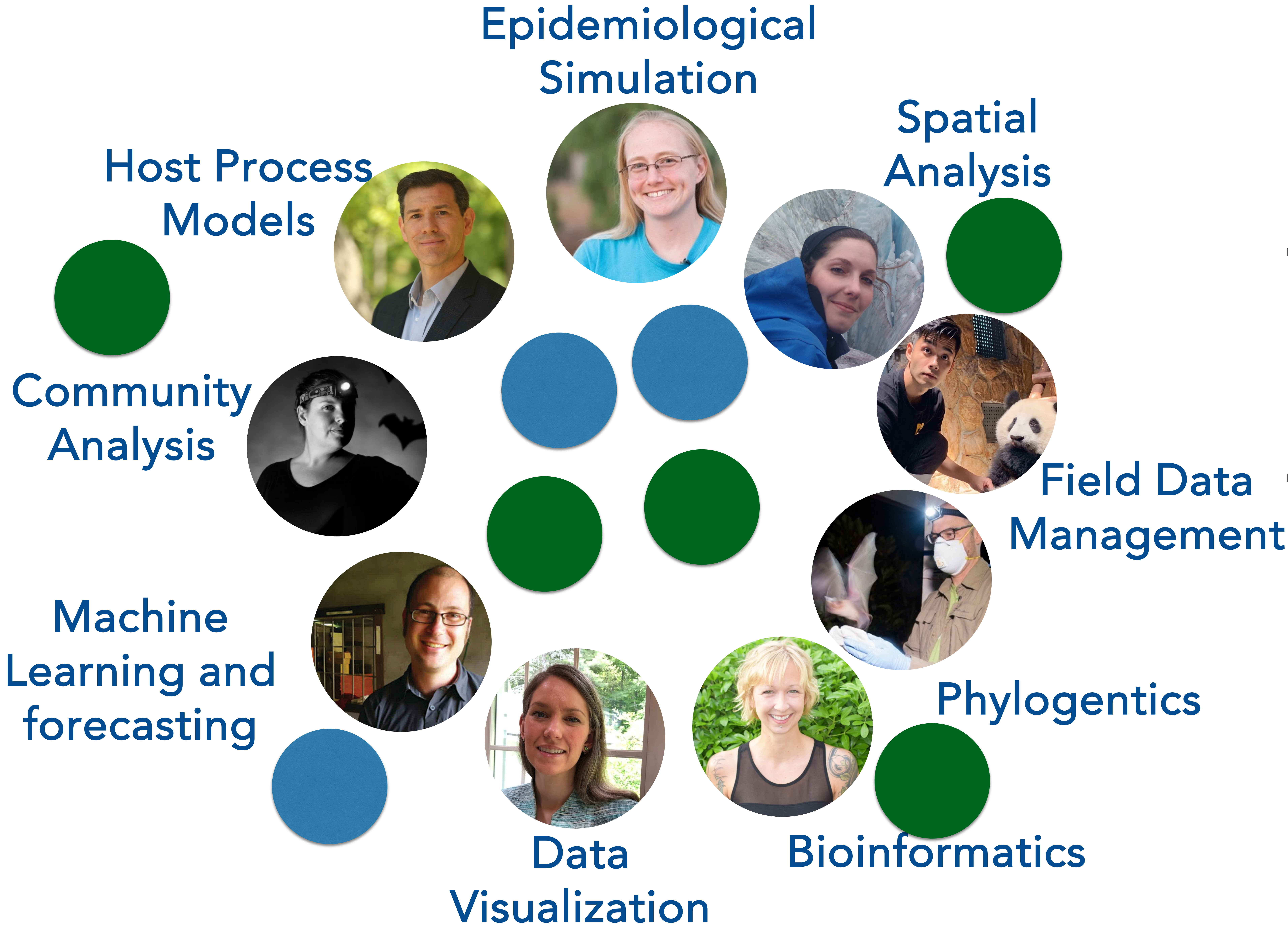
What Problems Arise at Team or Organization Scale?

- How do we collaborate efficiently within and across disciplines?
- How do we share best practices and maintain common standards of quality?
- How do we re-use information and work across projects simultaneously and in time?
- How do we share work while protecting privacy and security?



EcoHealth Alliance

- An interdisciplinary group of collaborators working on common scientific, conservation, and public health issues
- 50-60 staff, most scientific



- Data science needs and expertise live in every research group and project
- External collaborators play major roles in most projects

A Variety of Project and Output Types

- Data collection and QA processes for laboratory and field research
- High-compute machine learning, simulation, and genomics
- Mixed-methods analysis
- Scientific publications, data and figures
- Policy-relevant visualizations and syntheses
- “Live” dashboards and data feeds

A Common Workflow Language

(Not THE "Common Workflow Language")

- {targets} is an R-based build system for managing workflows and caching large computations
- Imposes and organizational framework on complex, multi-part projects
- Quite adaptable and scalable across different project types
- Works well on local, cloud or HPC
- Usually used with {renv} for dependency management, GitHub Actions for automation

A Common Workflow Language

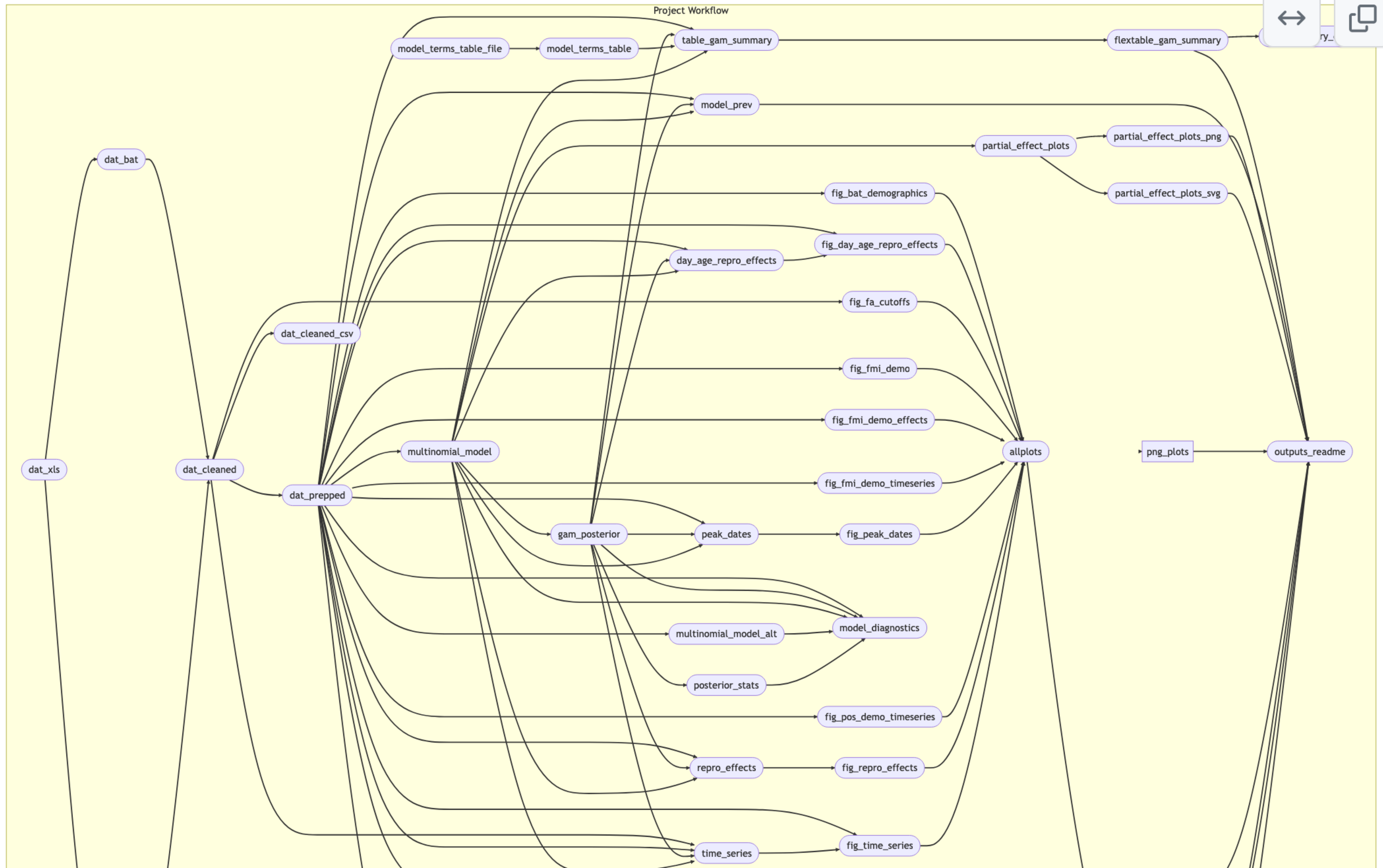
README.md

sabrenet-rousettus-dynamics

DOI 10.5281/zenodo.7709716 | License (for code) MIT | License (for data) CC0 1.0 | License (for text) CC BY 4.0

This repository contains data and R code which are supplements to:

Viral maintenance and excretion dynamics of coronaviruses within an Egyptian rousette fruit bat maternal colony – considerations for spillover, by Marike Geldenhuys, Noam Ross, Muriel Dietrich, John L. de Vries, Marinda Mortlock, Jonathan H. Epstein, Jacqueline Weyer, Janusz T. Paweska, and Wanda Markotter. *Scientific Reports* 13, 15829 (2023). https://doi.org/10.1038/s41598-023-42938-w

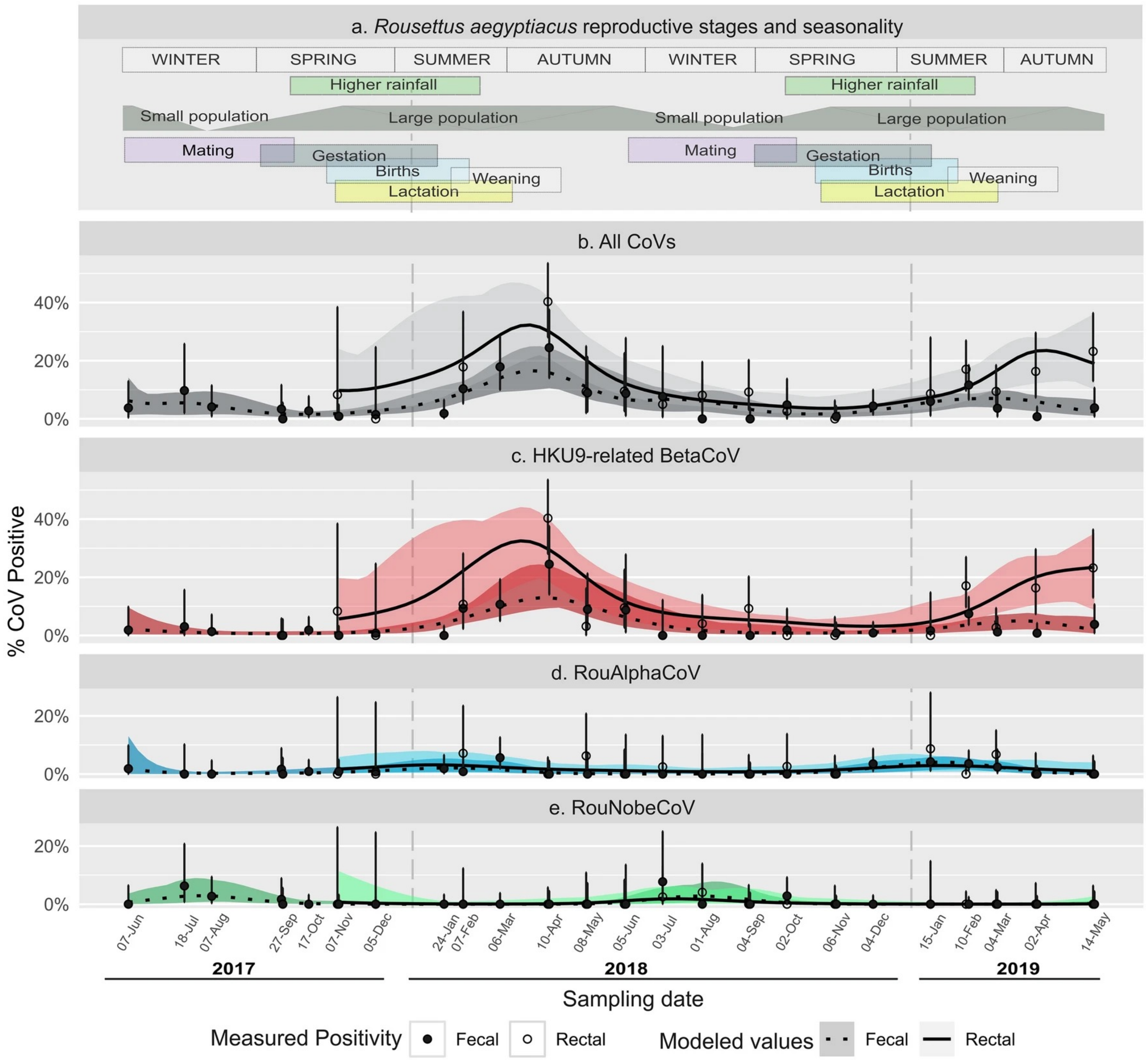


Article | Open access | Published: 22 September 2023

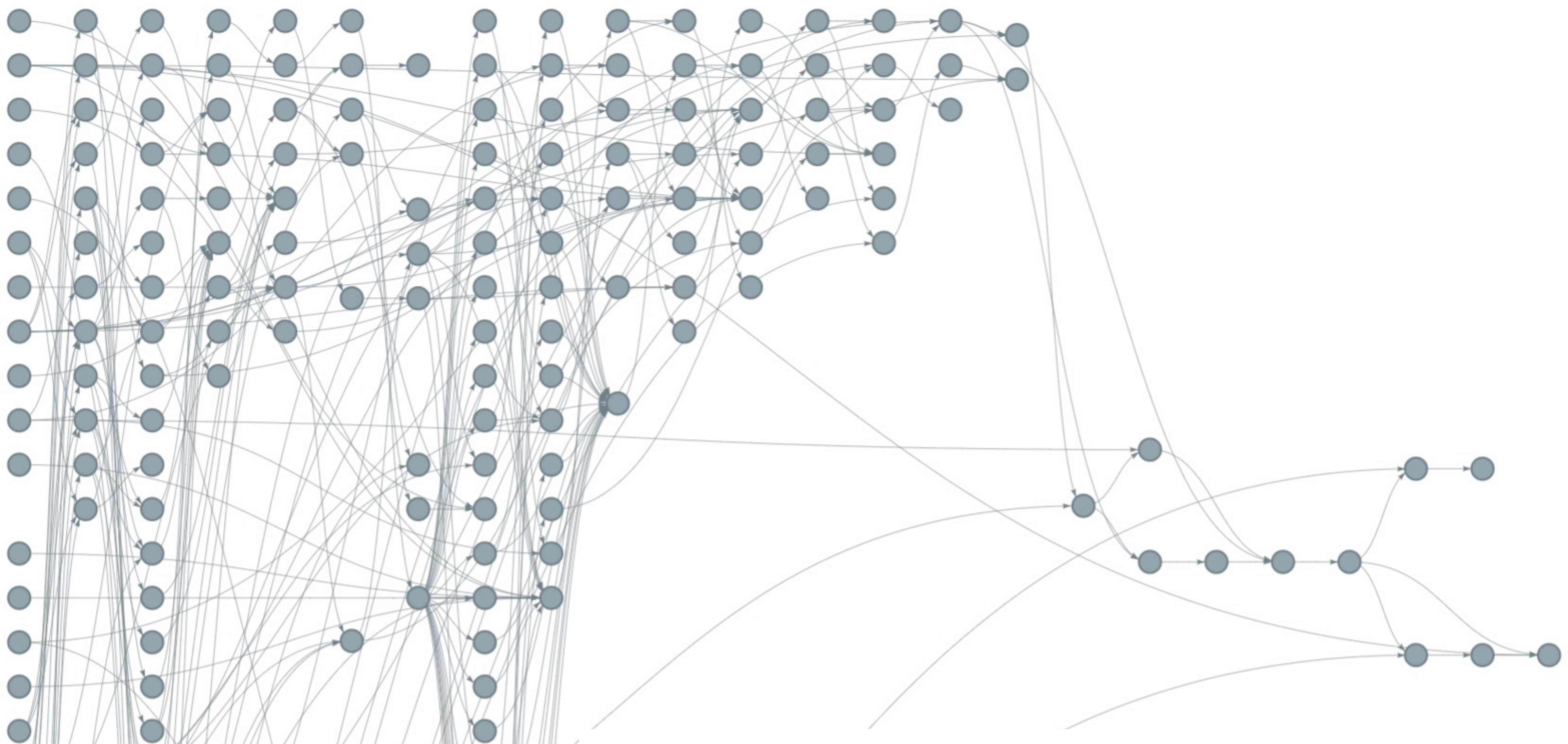
Viral maintenance and excretion dynamics of coronaviruses within an Egyptian rousette fruit bat maternal colony: considerations for spillover

Marike Geldenhuys, Noam Ross, Muriel Dietrich, John L. de Vries, Marinda Mortlock, Jonathan H. Epstein, Jacqueline Weyer, Janusz T. Paweska & Wanda Markotter

Scientific Reports **13**, Article number: 15829 (2023) | Cite this article

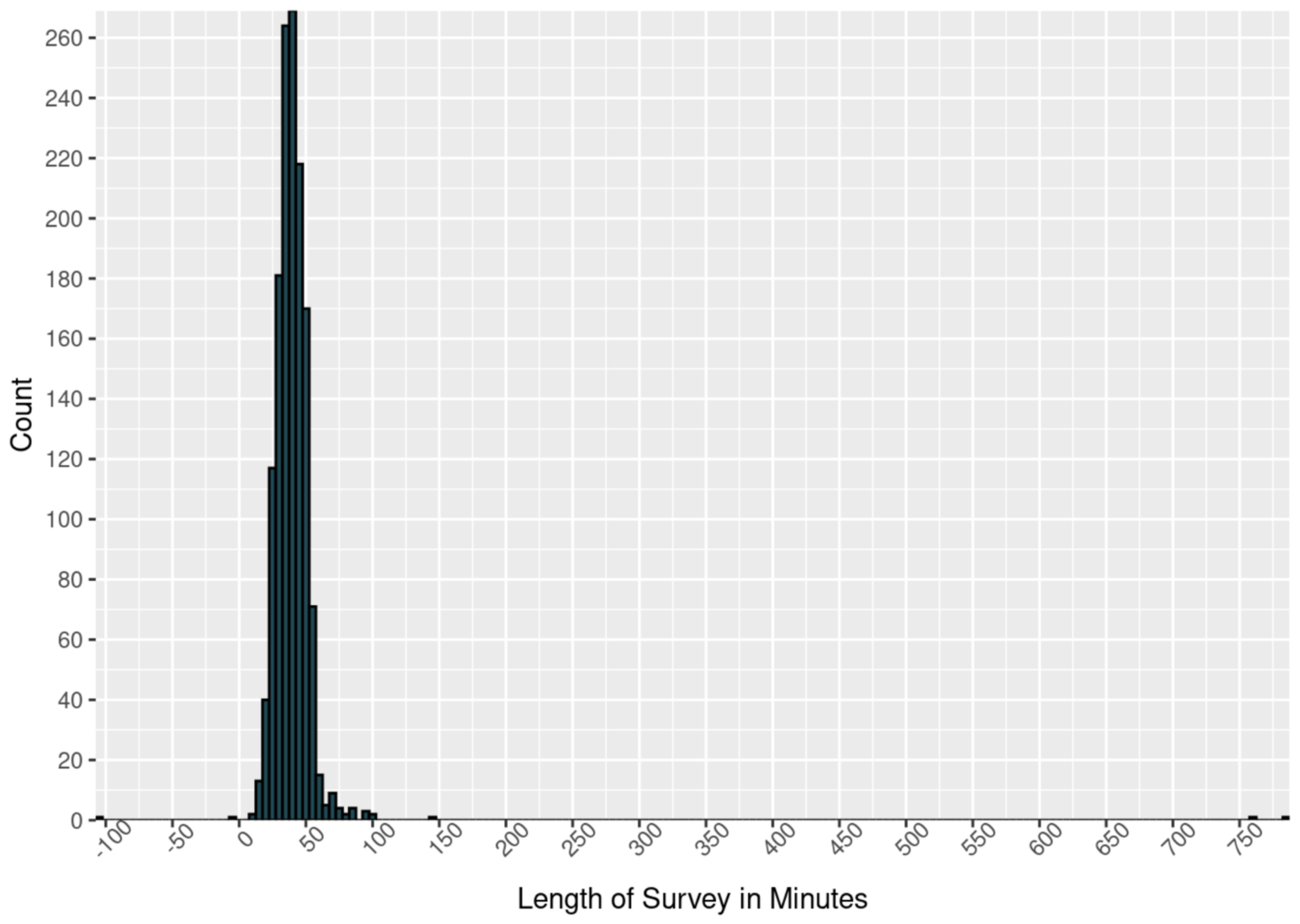


A Common Workflow Language



Survey Evaluation - II. Survey Lengths

Below is a histogram displaying the range of survey lengths. This is to assess how long it takes to complete the questionnaire.

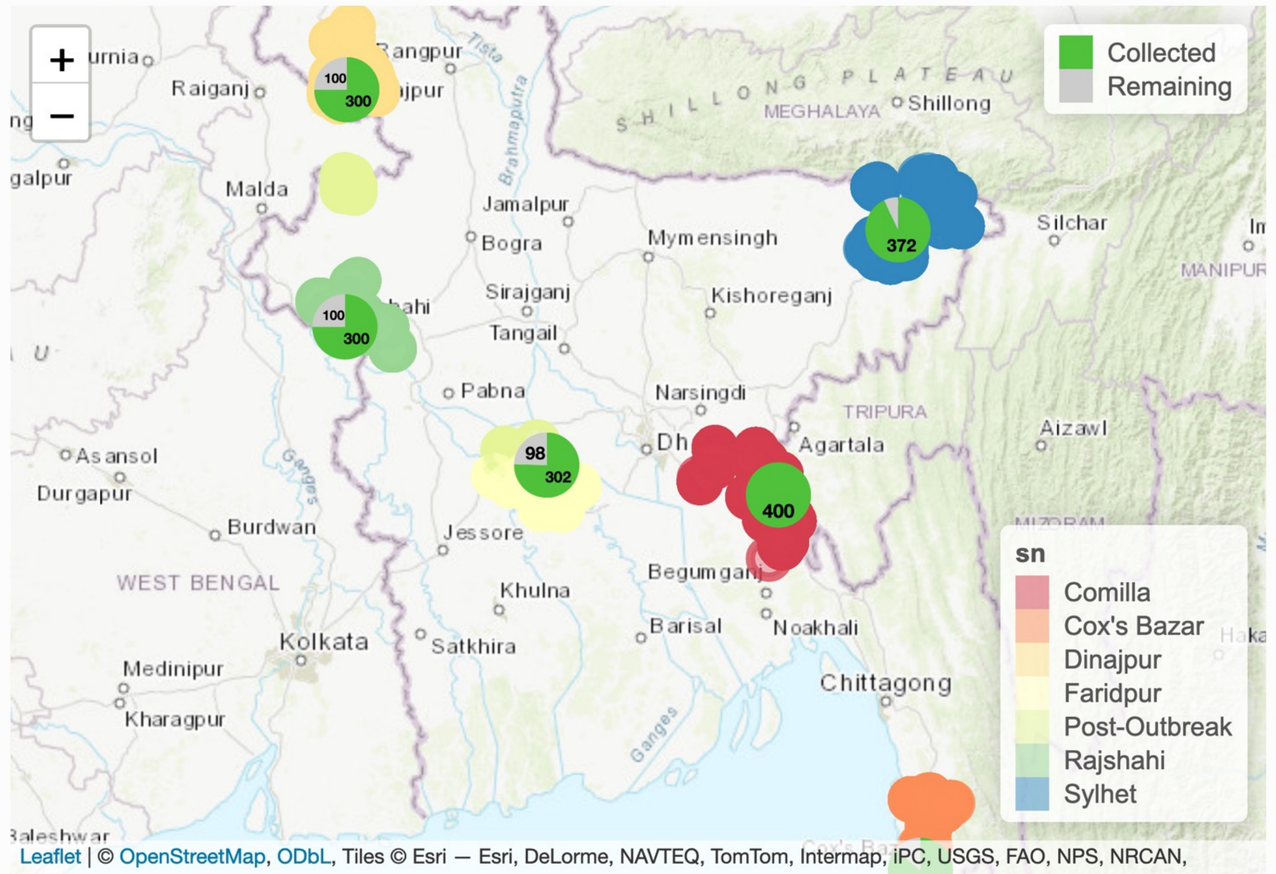


Data Summary - I. Number Surveys Collected

Survey Numbers | Interactive Map

The interactive map below displays the location of each survey as a color-coded dot. If you hold your cursor over the dots it will display the associated participant ID. When you click on the pie charts it displays the number of surveys collected for that site. "Other" (post-outbreak surveys without pre-specified site) surveys are shown, but there is no pie chart on the map.

Show

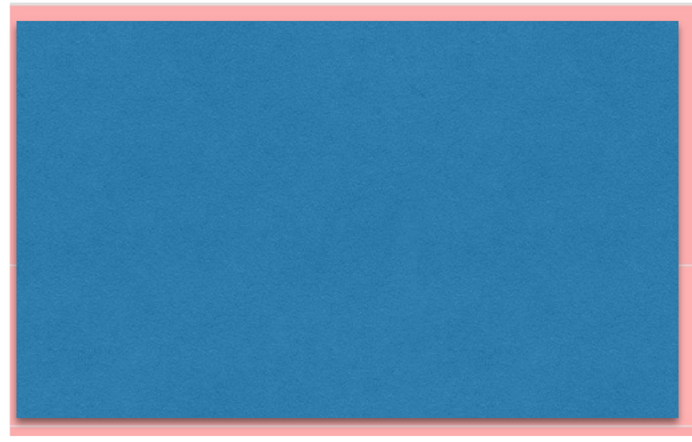


Leaflet | © OpenStreetMap, ODbL, Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

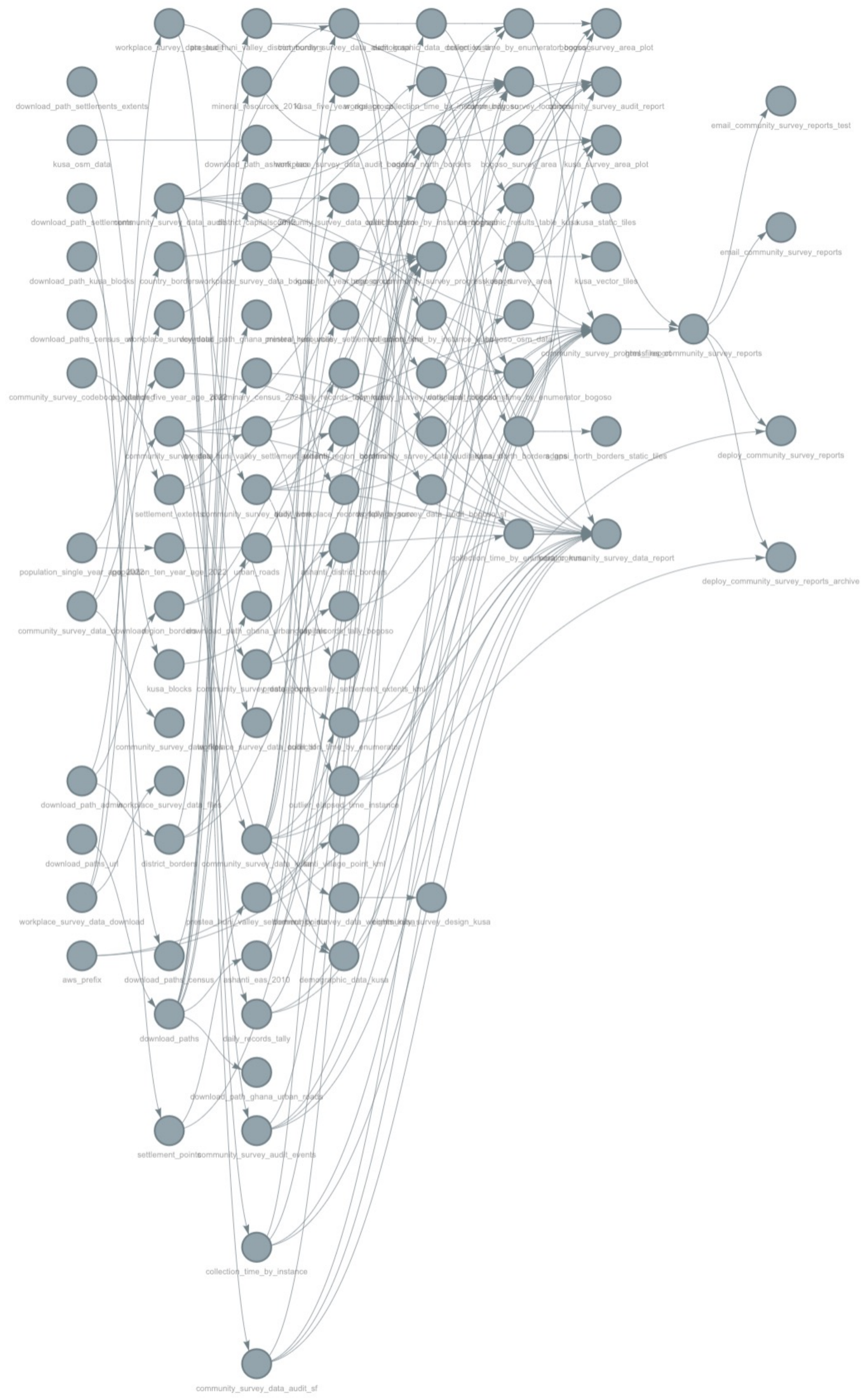
Plates Recorded	Issue Types	Records Added
0 (0%)	7	NA (NA)

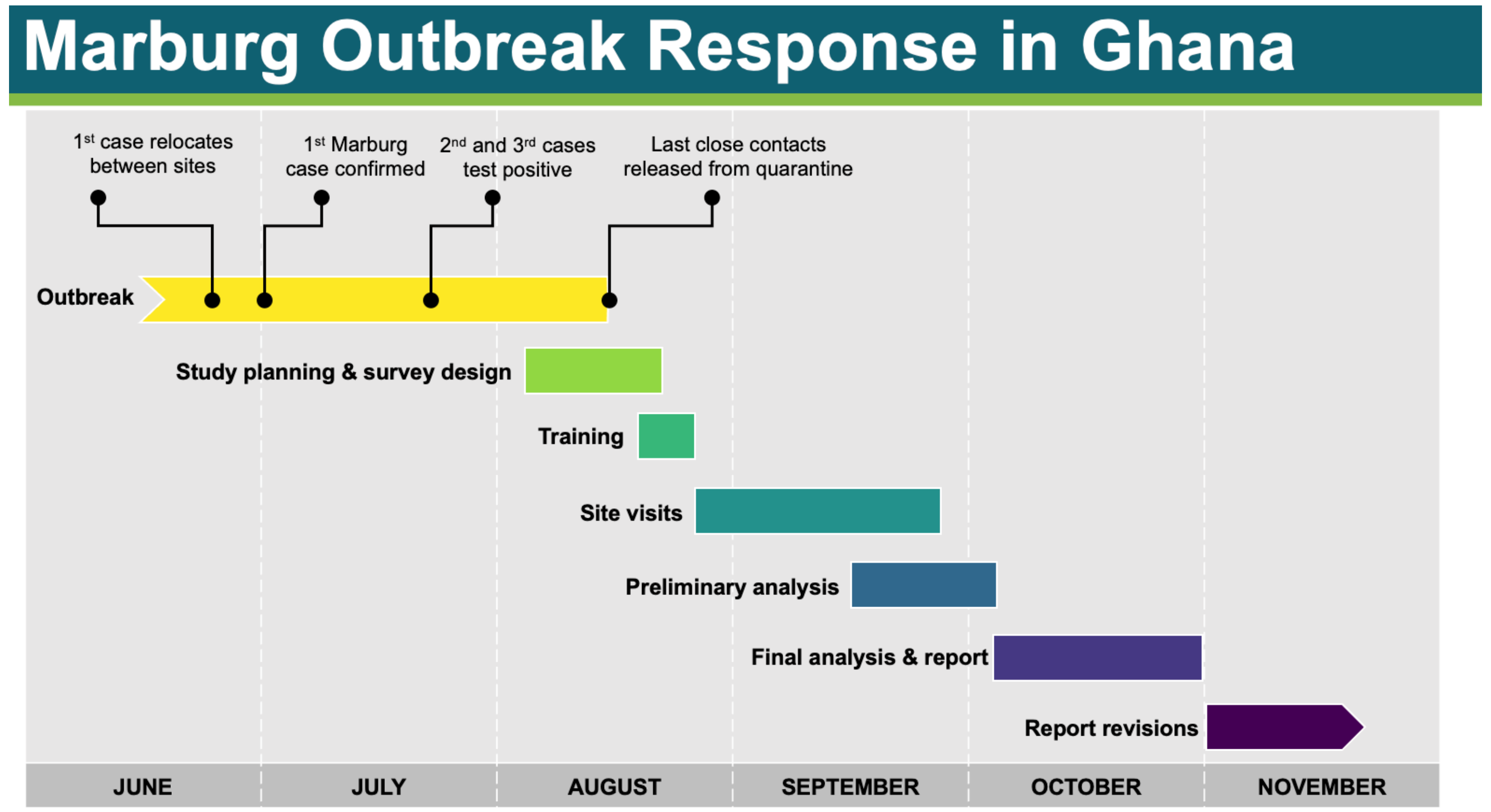
Show

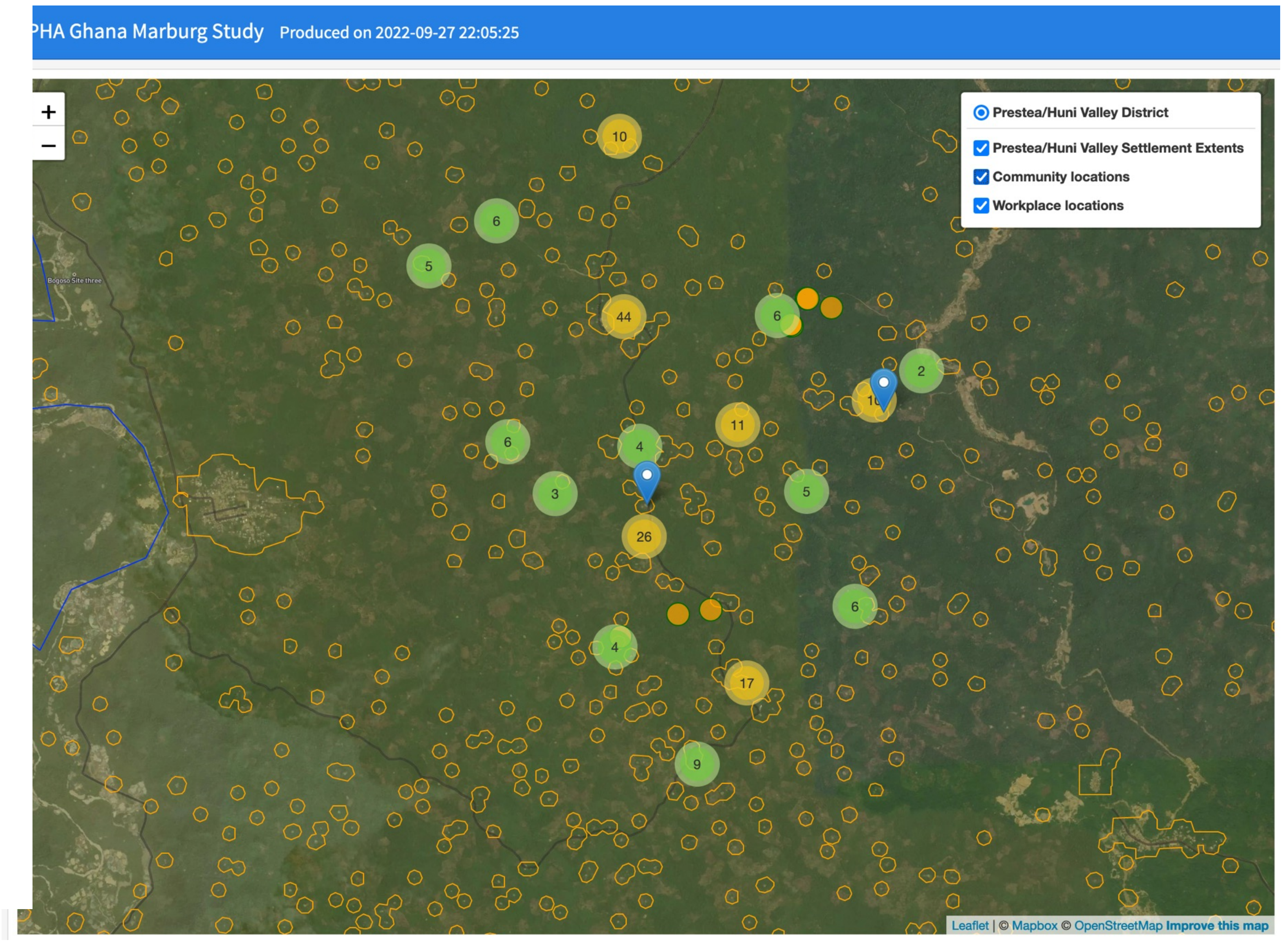
id	RT-PCR Plate ID	Issues	Recorded	createdTime
		RT-PCR Plate Layout: Excel extract does not have expected number of columns, RT-PCR Plate Layout: Discrepancy in the number of wells with samples , Specimen ID: Specimen id missing from ref table	FALSE	2023-04-26T06:19:23.000Z
		RT-PCR Plate Layout: Airtable and Excel RT-PCR Plate ID do not match, Specimen ID: Specimen id missing from ref table	FALSE	2023-10-11T06:32:32.000Z



A Common Workflow Language





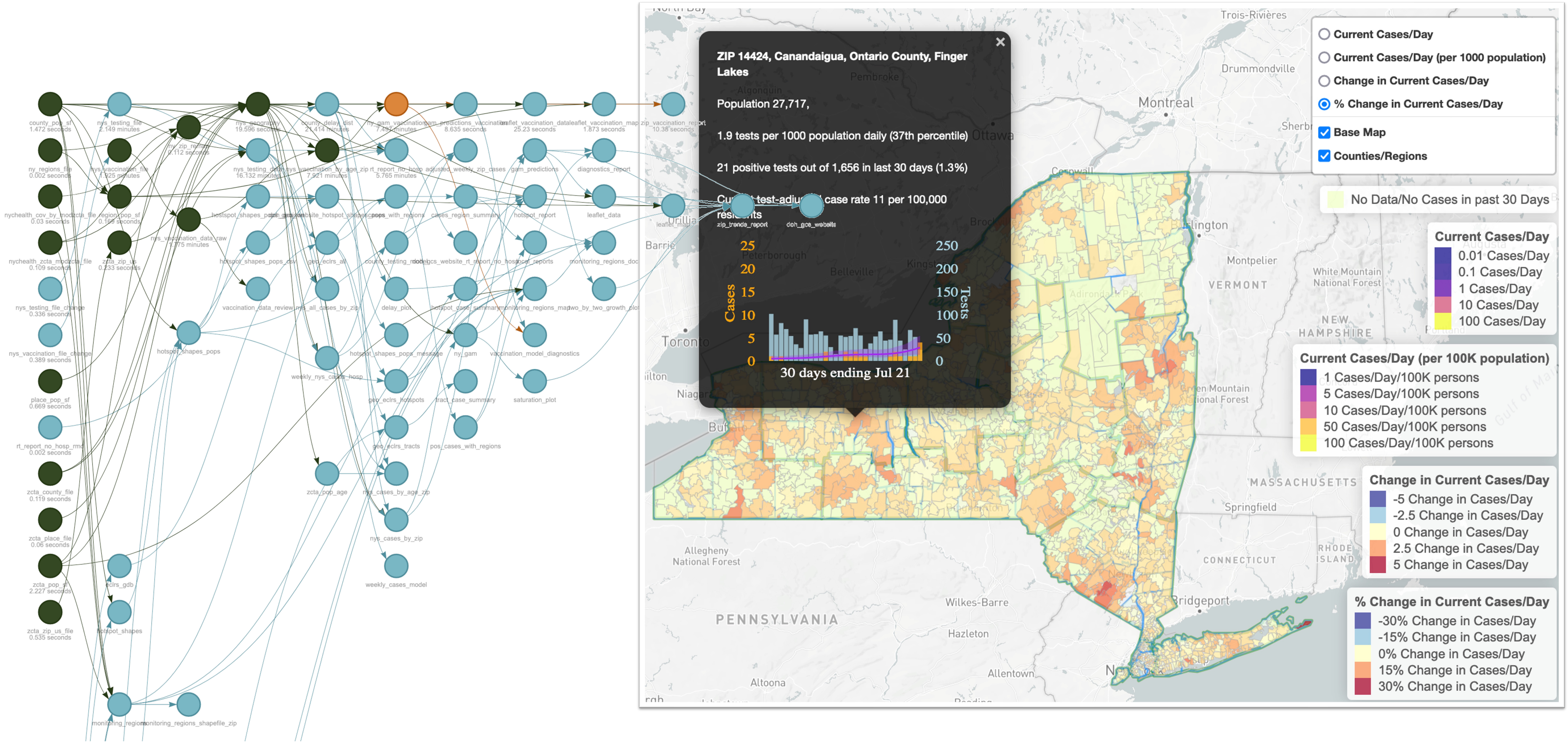


Number of records completed per day by enumerator as of 2022-09-27

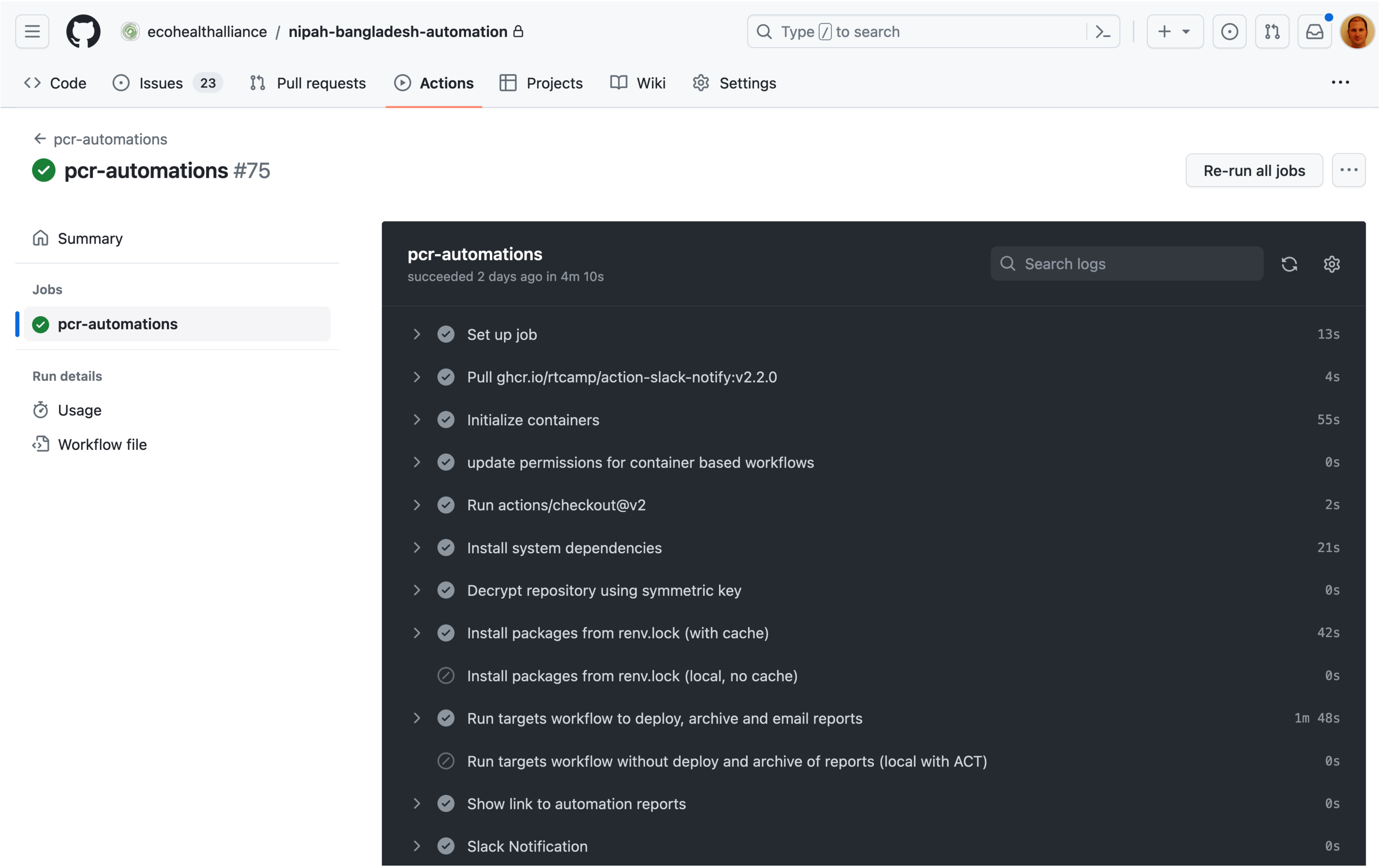
	ASA	MYZ	RDS	SOB	Total
2022-09-15	7	7	0	0	14
2022-09-16	13	9	2	4	28
2022-09-17	16	14	9	7	46
2022-09-18	15	9	10	8	42
2022-09-20	20	14	11	12	57
2022-09-21	15	12	11	10	48
2022-09-22	10	11	3	3	27
2022-09-23	12	10	5	9	36
2022-09-24	12	12	6	6	36
2022-09-25	9	9	0	7	25
2022-09-26	17	10	0	11	38
Total	146	117	57	77	397

A total of **397** records have been completed as of **2022-09-27**.

A Common Workflow Language

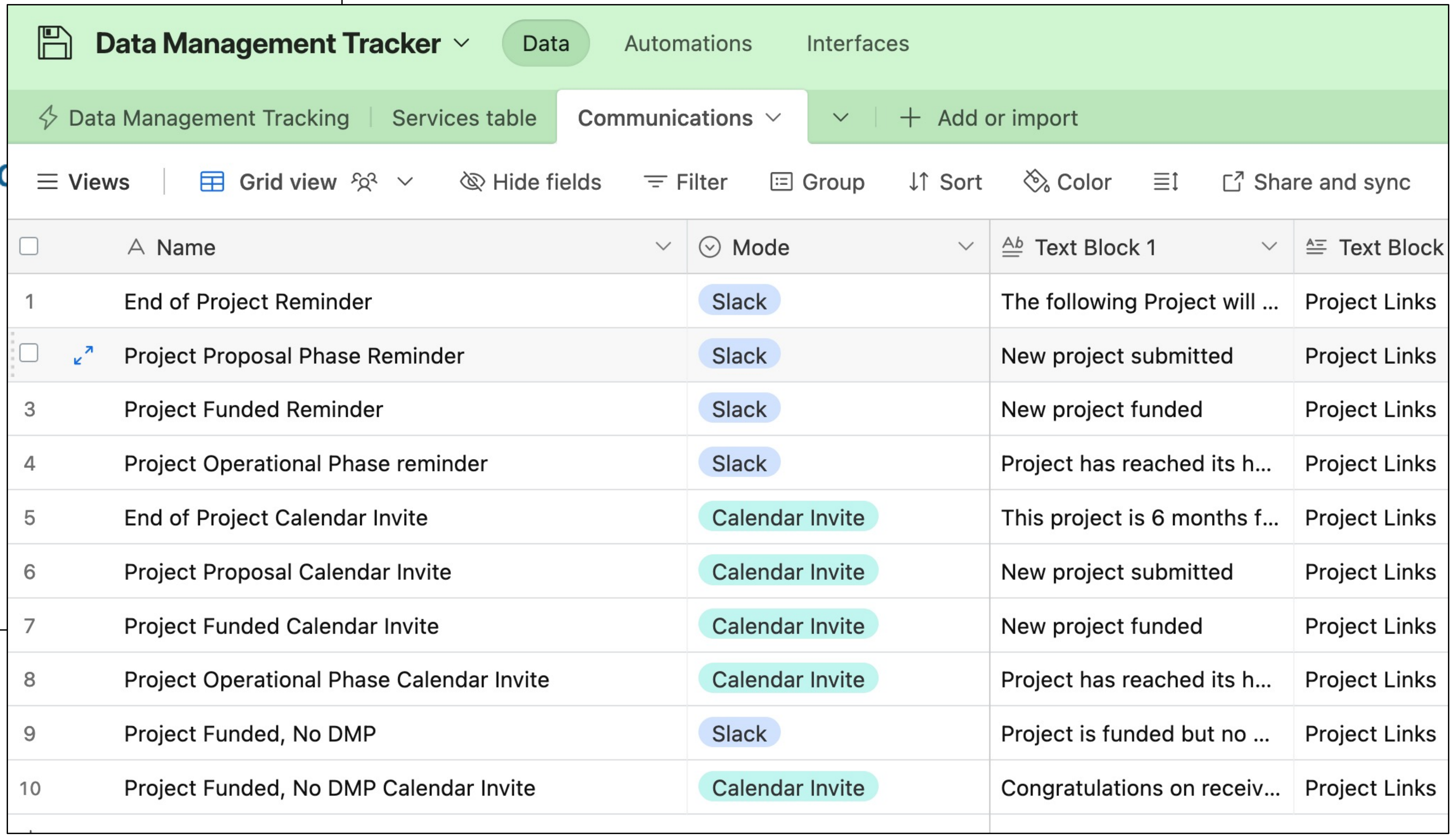


Process Automation – GitHub Actions



Managing Privacy/Security/Openness Trade-offs

- A structured workflow and tracking for data-sharing plans is is essential



	Name	Mode	Text Block 1	Text Block
1	End of Project Reminder	Slack	The following Project will ...	Project Links
	Project Proposal Phase Reminder	Slack	New project submitted	Project Links
3	Project Funded Reminder	Slack	New project funded	Project Links
4	Project Operational Phase reminder	Slack	Project has reached its h...	Project Links
5	End of Project Calendar Invite	Calendar Invite	This project is 6 months f...	Project Links
6	Project Proposal Calendar Invite	Calendar Invite	New project submitted	Project Links
7	Project Funded Calendar Invite	Calendar Invite	New project funded	Project Links
8	Project Operational Phase Calendar Invite	Calendar Invite	Project has reached its h...	Project Links
9	Project Funded, No DMP	Slack	Project is funded but no ...	Project Links
10	Project Funded, No DMP Calendar Invite	Calendar Invite	Congratulations on receiv...	Project Links

Managing Privacy/Security/Openness Trade-offs

README.md

git-crypt - transparent file encryption in git

git-crypt enables transparent encryption and decryption of files in a git repository. Files which you choose to protect are encrypted when committed, and decrypted when checked out. git-crypt lets you freely share a repository containing a mix of public and private content. git-crypt gracefully degrades, so developers without the secret key can still clone and commit to a repository with encrypted files. This lets you store your secret material (such as keys or passwords) in the same repository as your code, without requiring you to lock down your entire repository.

git-crypt was written by Andrew Ayer (agwa@andrewayer.name). For more information, see https://www.agwa.name/projects/git-crypt.

- Tools that allow for private and public phases of work are extremely helpful

- git-crypt is a handy tool for encrypting *parts* of git repositories in a platform-independent way

Data Platforms for all Disciplines

No-Code Interfaces, Structured Data and API back-ends

OPEN DATA KIT

- Tablet-based field data collection platform
- Designed for robustness and offline environments
- Open source, self-run on cloud servers (service available)

Airtable

- Spreadsheet-like front-end, database back-end
- Great for bulk data entry
- Provides a common language and platform data for many stakeholders and processes across the organization
- $$$!

Data Platforms for all Disciplines

OPEN DATA KIT

DTRA Jordan

Overview | Project Roles | App Users | Form Access | Settings

About Projects

Projects let you group related Forms and Users. Web Users can be given Roles that let them perform certain actions within this Project, including using a web browser to fill out Forms. App Users for this Project can only see Forms in this Project which they have access to.

For more information, please see this help article. If you have any feedback, please visit this forum thread.

Right Now

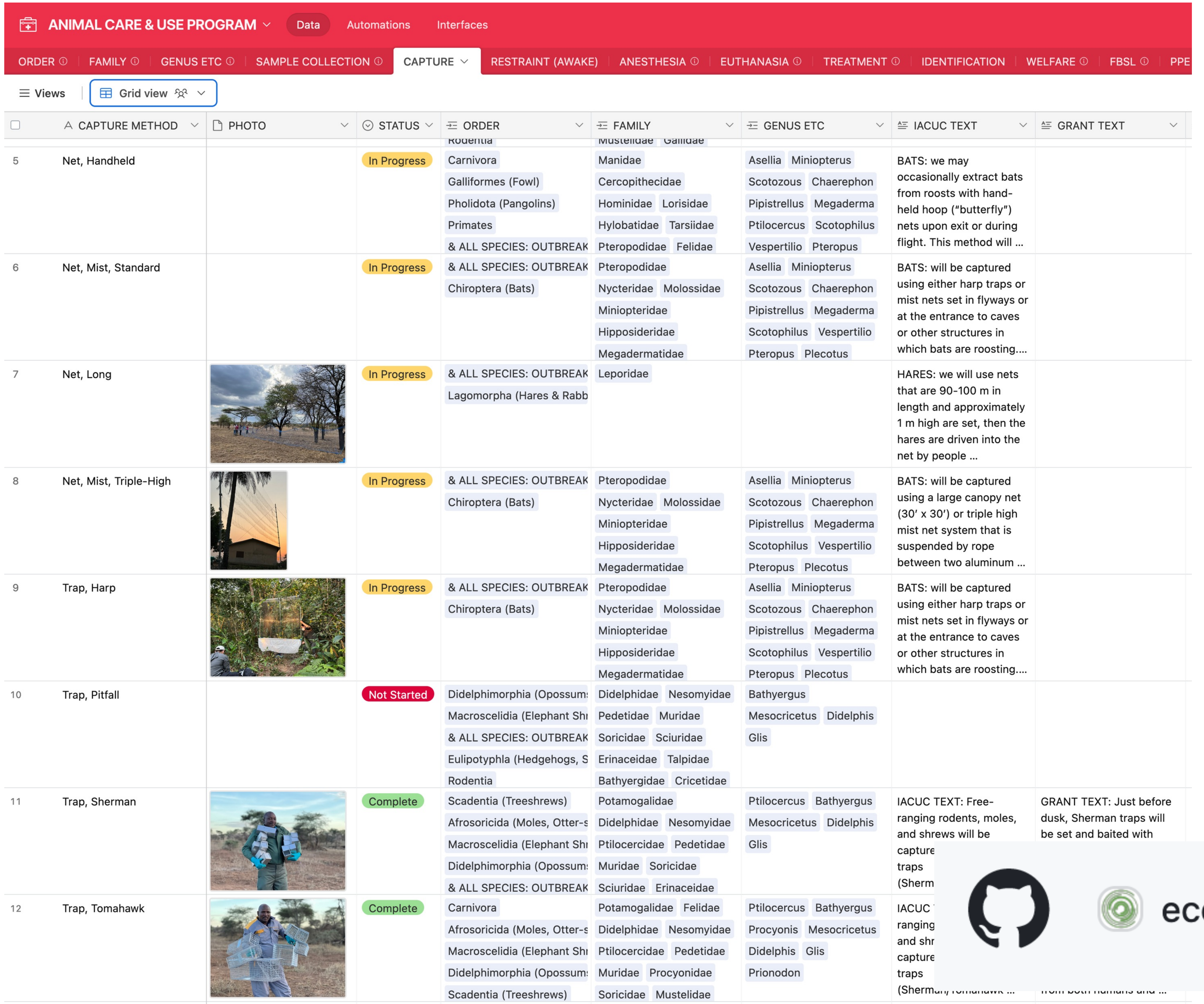
11 ›
App Users who can use a data collection client to download and submit Form data to this Project.

7 ›
Forms which can be downloaded and given as surveys on mobile clients.

Name	ID and Version	Submissions	Actions
Animal_History_Form ›	Animal_History_Form 20220202	279 Submissions › (last 2023/10/25 11:31)	Preview
Follow_up_questionnaire ›	Follow_up_questionnaire 20230405	2,209 Submissions › (last Monday 04:34)	Preview
Follow_up_questionnaire_draft ›	Follow_up_questionnaire_draft	1 Submission › (last 2022/04/26 14:34)	Preview
HQ_Test ›	HQ_Test 20220207	10 Submissions › (last 2022/02/07 10:17)	Preview
Human_Questionnaire ›	Human_Questionnaire 20220207	499 Submissions › (last 2022/10/01 02:42)	Preview
Site_Interface_Characterization_Form ›	Site_Interface_Characterization_Form 20220511	276 Submissions › (last 2023/10/25 11:31)	Preview
Workshop Demonstration of ODK ›	workshop_survey 20220207	31 Submissions › (last 2022/10/18 04:31)	Preview

ropensci / ruODK

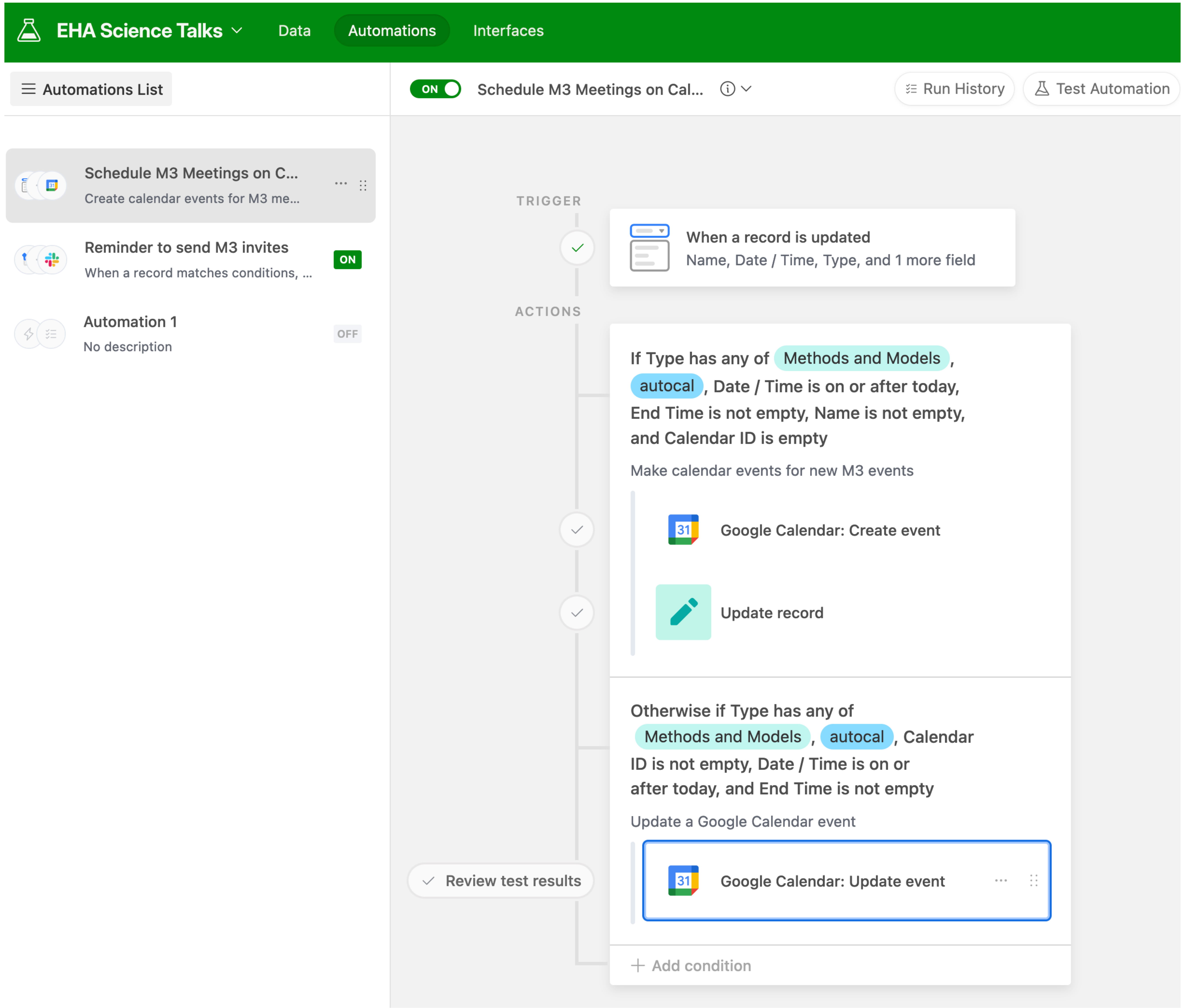
Data Platforms for all Disciplines



Airtable

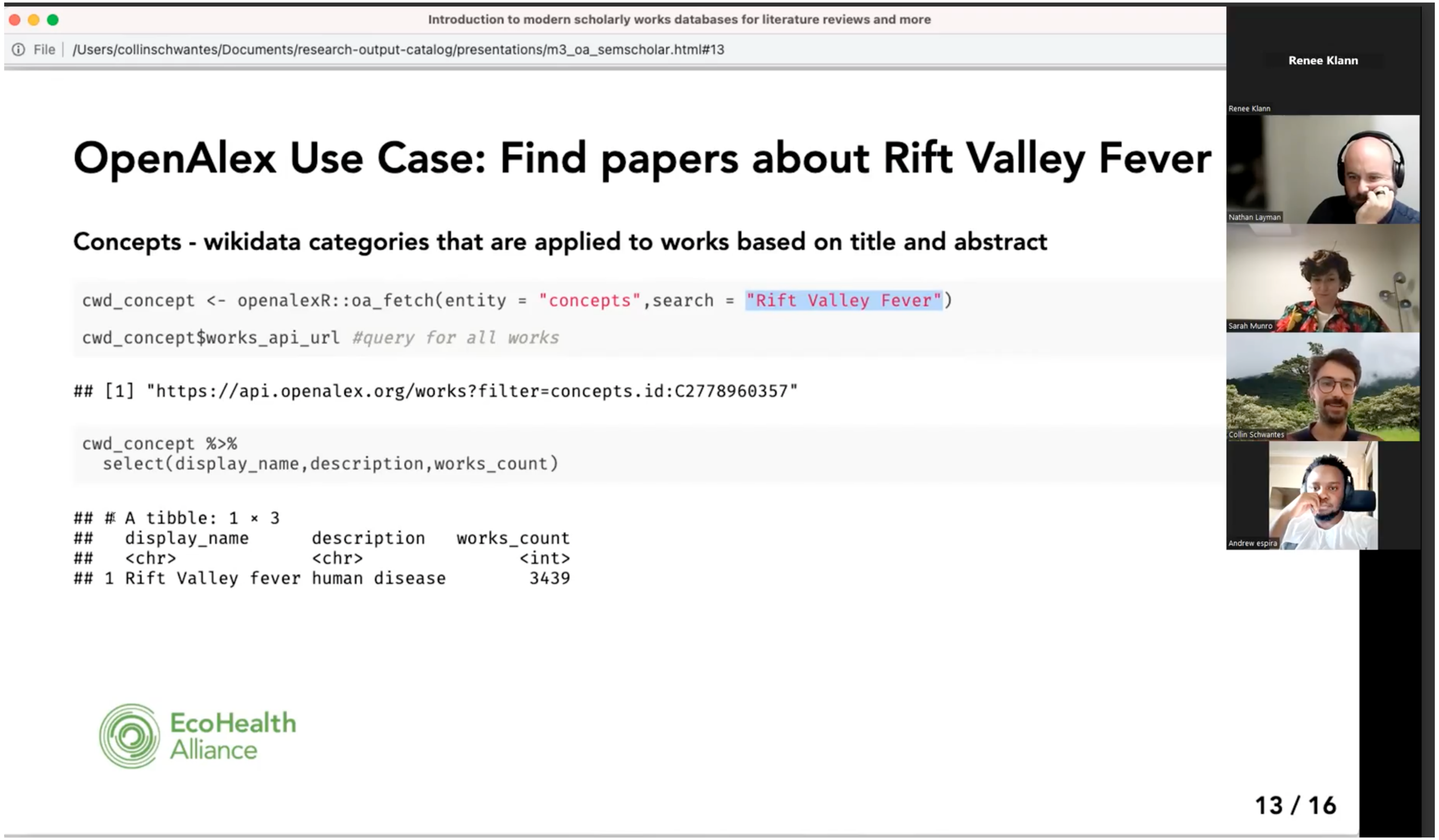
ecohealthalliance / airtabler

Process Automation



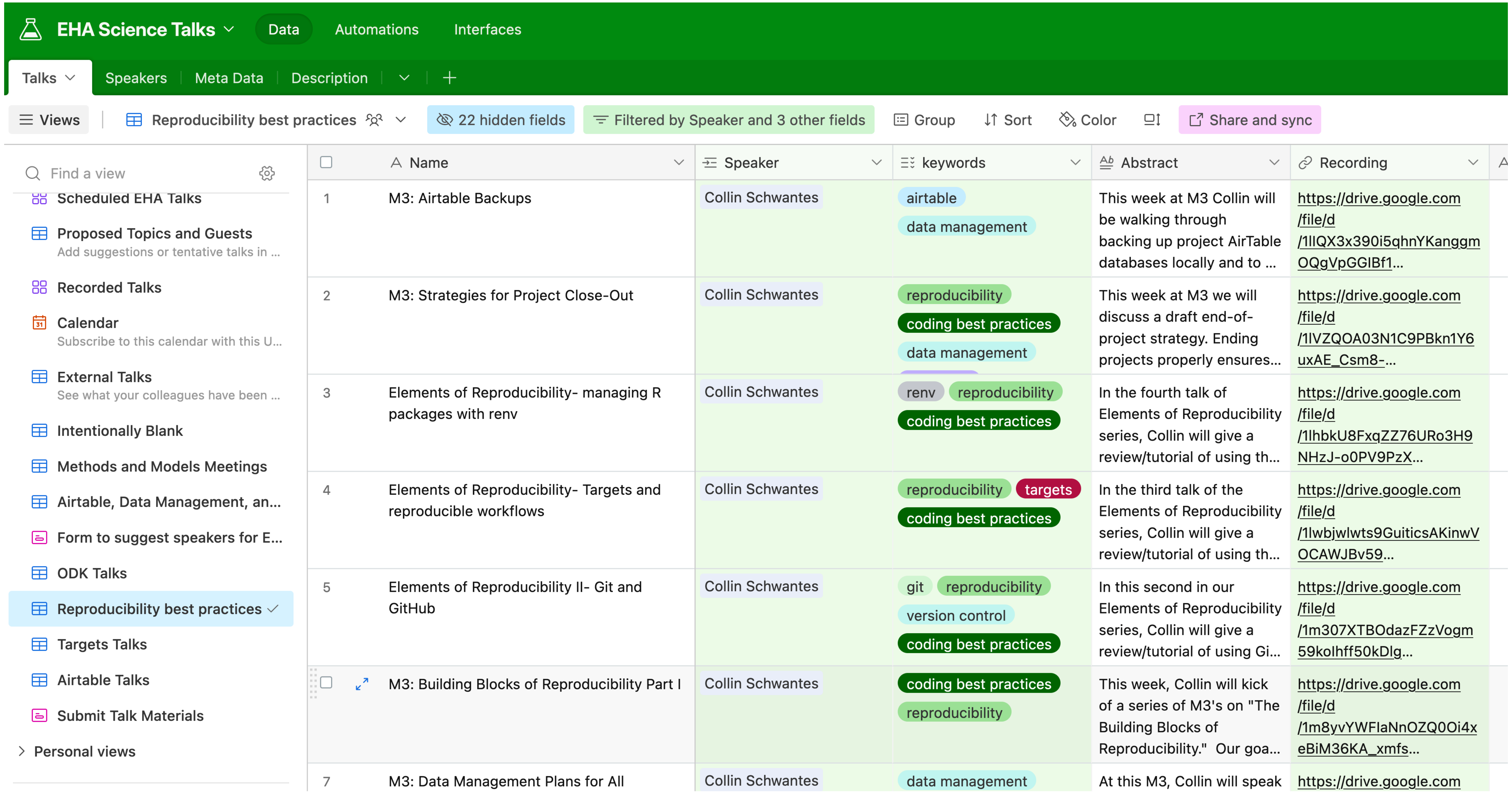
Airtable

Cross-Training Across Teams



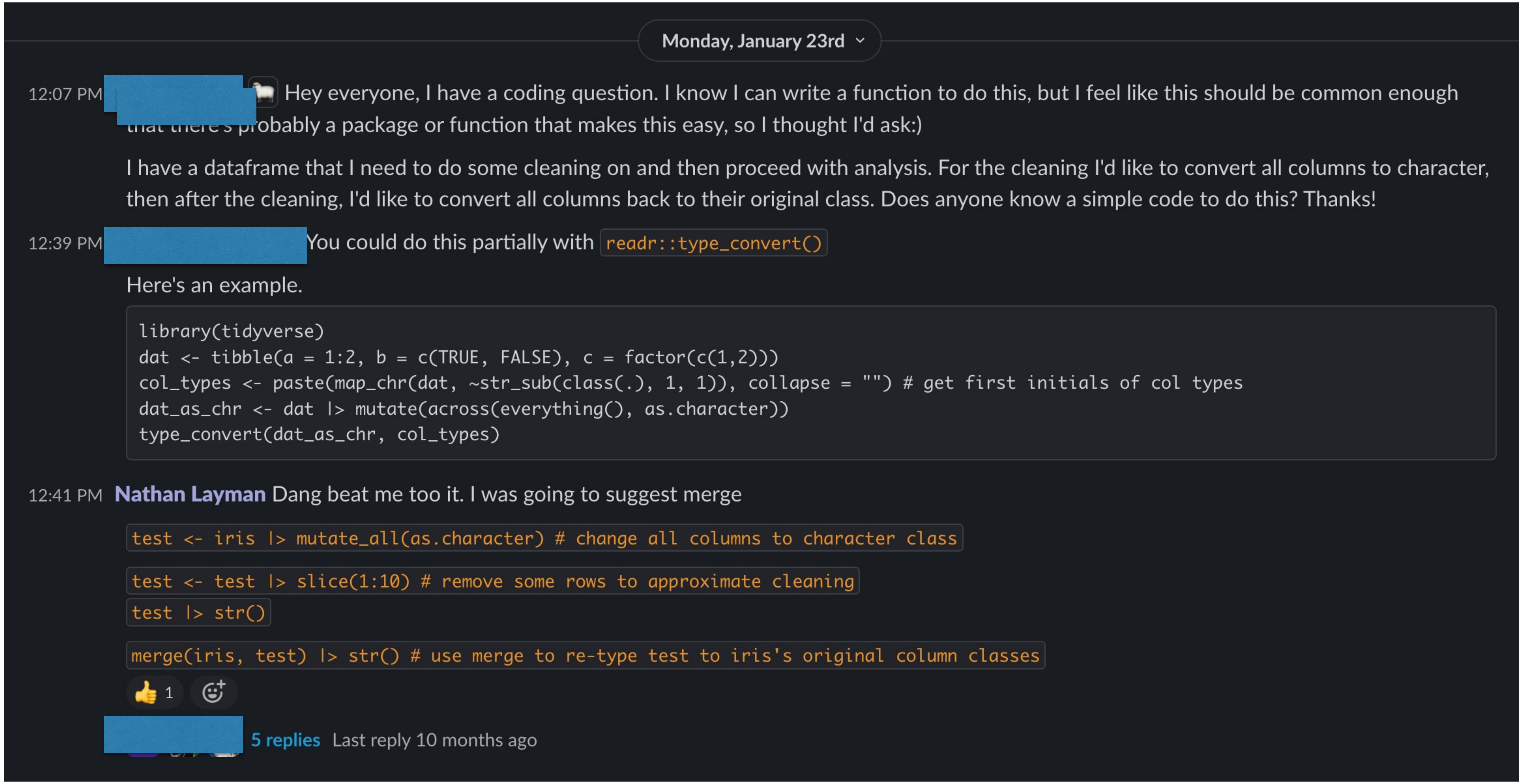
- Regular "Methods and Models Meetings" focus on sharing new approaches and best practices across teams
- Also brainstorming and feedback sessions for in-development ideas
- Open to external collaborators

Cross-Training Across Teams



- Recorded talks live in a searchable catalog of topics

Cross-Training Across Teams



- Dedicated chat channels for cross team support topics:
- #data-sci-discuss
- #eha-servers
- #grants-q-and-a…

Common Resources of Expertise

Infrastructure Engineer
HPC Management
Cloud resource provisioning
Field system testing and training

Data Librarian
Project data system design
Data lifecycle management
Reproducibility support and training

- Some data science expertise is centralized in cross-team experts
- As with finance, grant compliance, etc…

Onboarding and Training

EcoHealth Alliance

Modeling & Analytics Handbook

EHA Modeling & Analytics Handbook

"These aren't rules, just some things that we figured out." – Michael Reno Harrel

Last edit 2023-09-12 by Collin Schwantes

Introduction

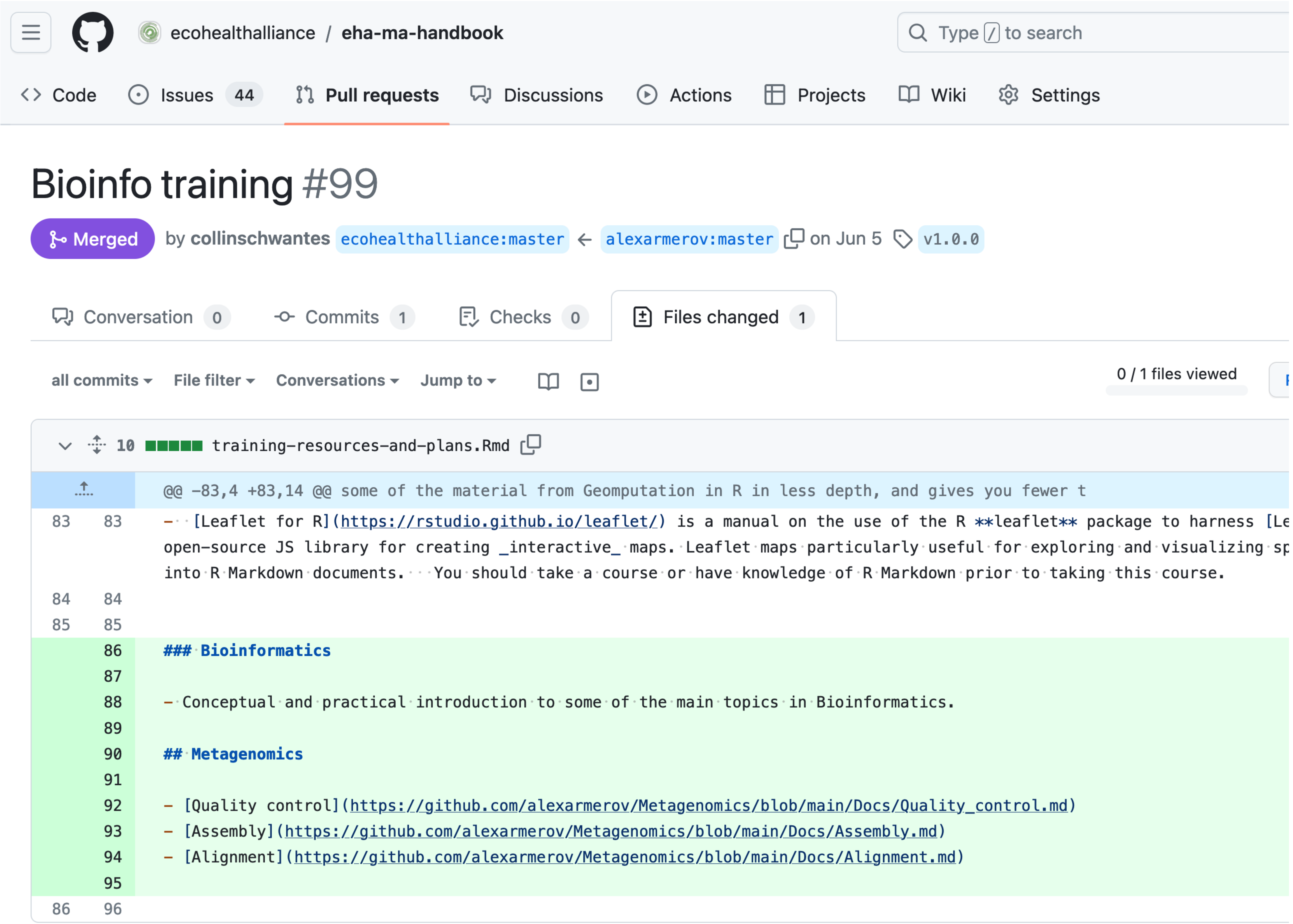
This handbook describes best practices and guidelines for project management, organization, modeling and and programming we aim for our on the EHA Modeling & Analytics team.

This is a living document. To make changes, just click the edit button () at the top of the page. It will take you to the source editor for the chapter on GitHub, where you can make edits and submit your changes. Be sure to commit major contributions to a new branch and open a pull request.

- An open handbook serves as a knowledge repository for important topics

ecohealthalliance.github.io/eha-ma-handbook/

Onboarding and Training



- New resources and updates are added to the handbook on an ongoing basis, often driven by seminars.

Onboarding and Training

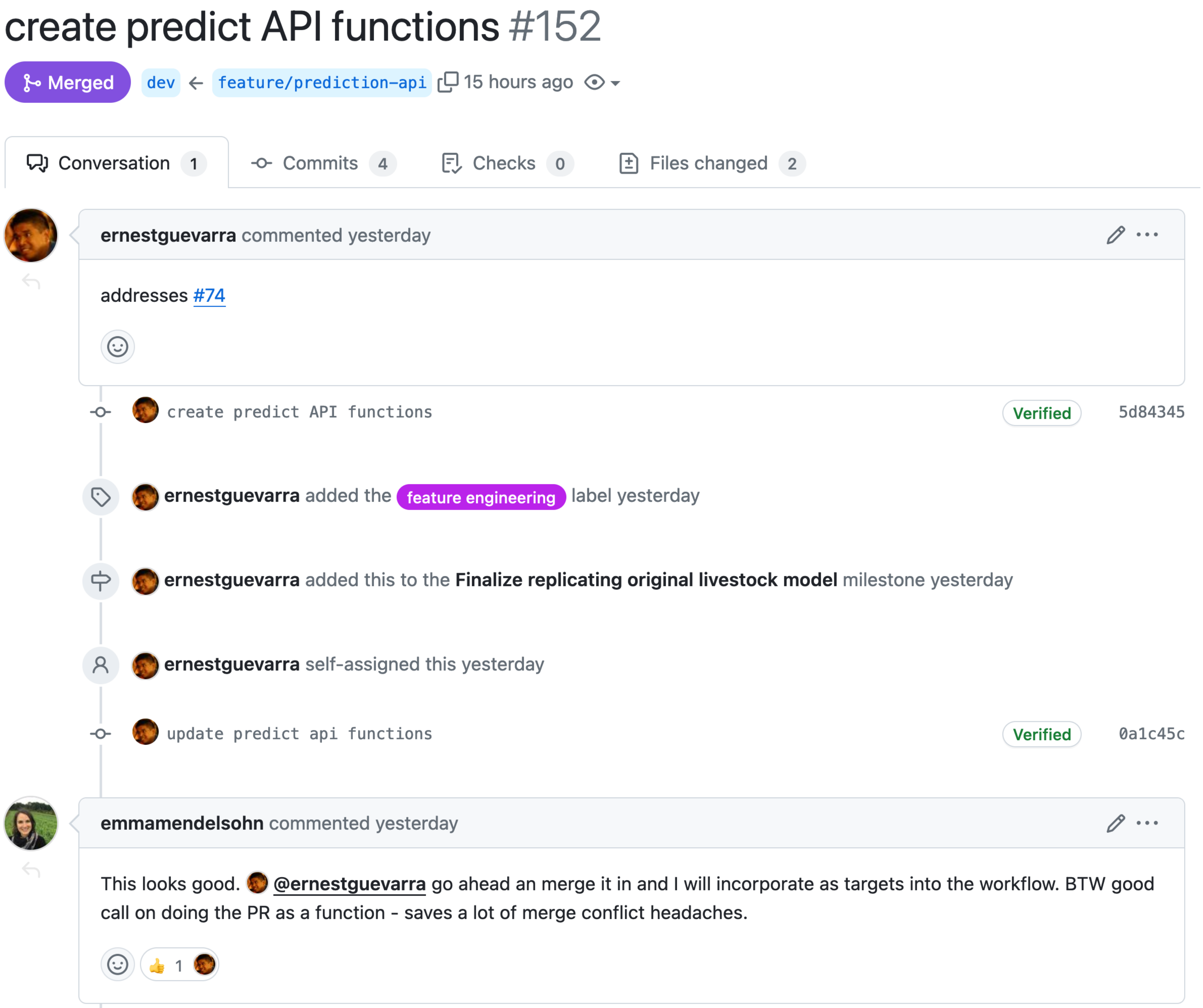
Learn about the organization and team

Over the first month

- Read up on the REPEL project
  - Our technical report from Phase I REPEL-technical-report.pdf
  - the proposal and work plan for Phase II Proposal, Technical and Management Approach.pdf
- Browse the dashboard of EHA projects to learn about the different work we are doing across the organization
- Learn about our EHA's departments and committees.
- We are starting to migrate shared resources to this All-Staff Resources AirTable and Google Drive Folder, so it should be the first place you look for things and, if important things are not there, ask that they be migrated.
- Learn about the tooling we use in our projects:
  - AirTable for shared data and workflow management
  - git and GitHub for code version management
  - targets for workflow
  - renv for dependency management
  - dolt for versioned data (and our nascent R client package)
- Get familiar with GitHub Project Boards and our approach to using them. We're using them most on the REPEL2 project, so you'll get an intro to this on Wednesday.
- Skim the Modeling and Analytics Manual
- Peruse and watch some recent talks in our science talks archive.
- Look at this folder of a few field and policy projects started recently to get a sense of some of the projects going on across EHA.

- New team members get an onboarding document directing them to these and other resources

Co-Creation Mechanisms



- Peer review of code in projects
- Co-work sessions
- Team stand-ups and problem-solving sessions

Thank You!

ross@ecohealthalliance.org
@noamross@ecoevo.social

EcoHealth Alliance

ecohealthalliance.org

ROpenSci

ropensci.org