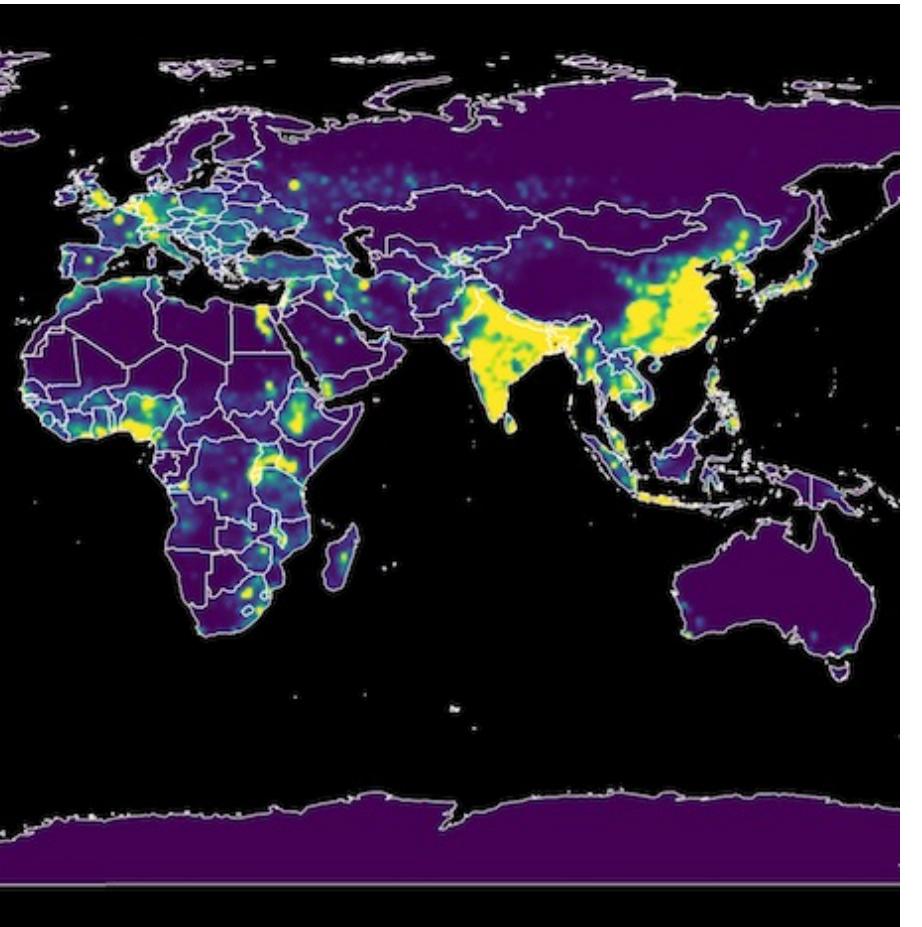
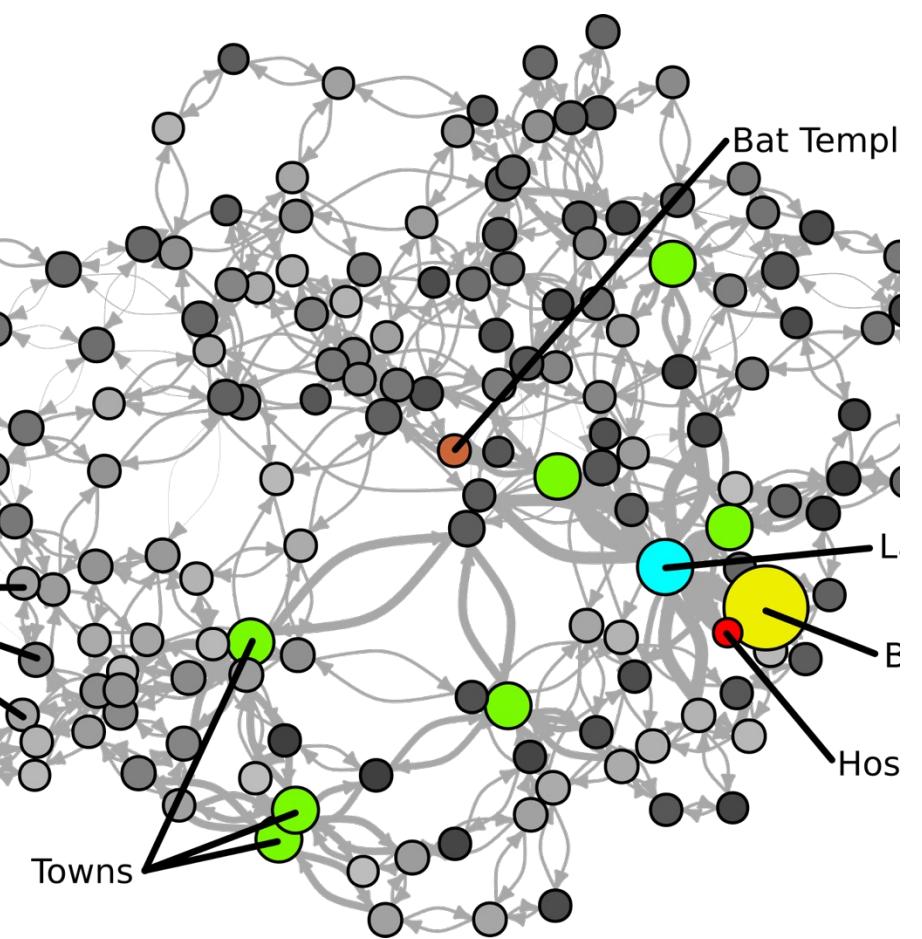


Tools and Teams for Reproducible Science with R

Noam Ross, EcoHealth Alliance

R User Group at the Harvard Data Science Initiative

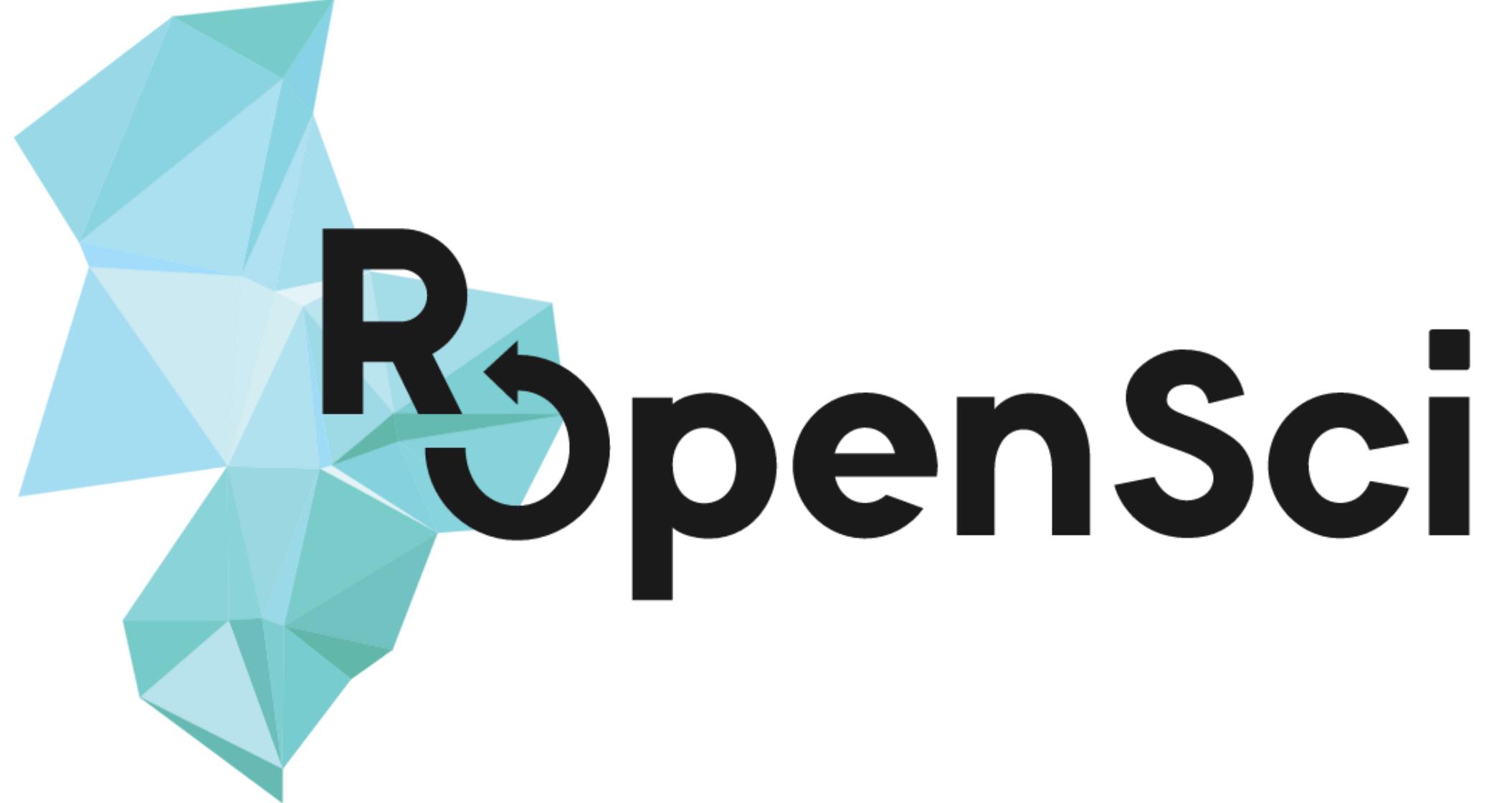
2021-07-22



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

- Overview
- Introduction
- Data management
- Software
- Collaboration
- Project organization
- Keeping track of changes
- Manuscripts
- What we left out

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.

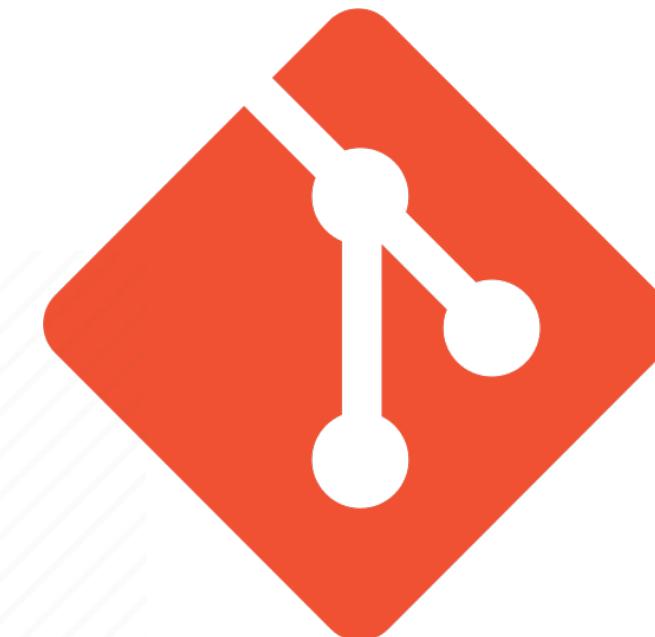
5. Keeping track of changes

- 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

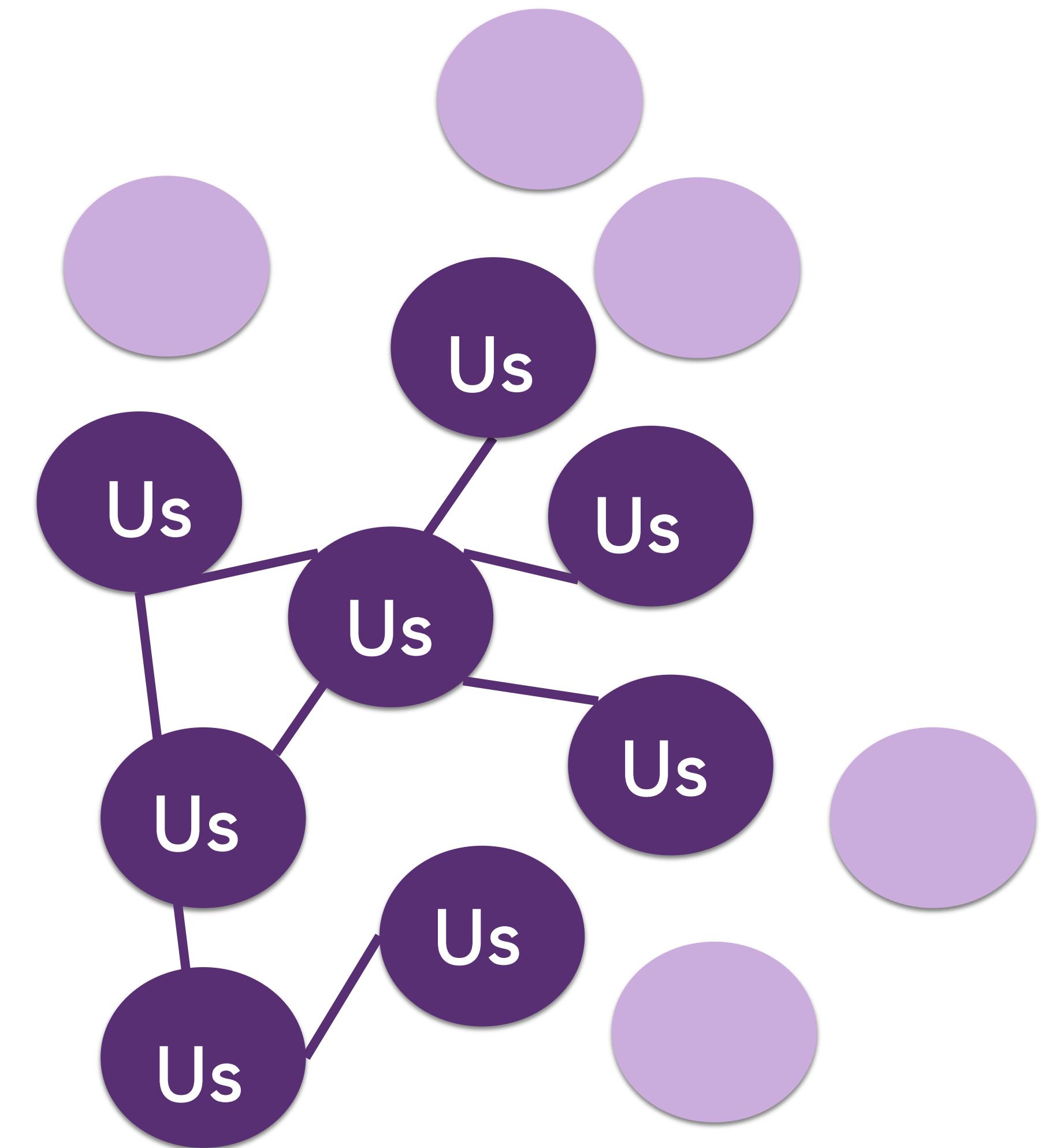
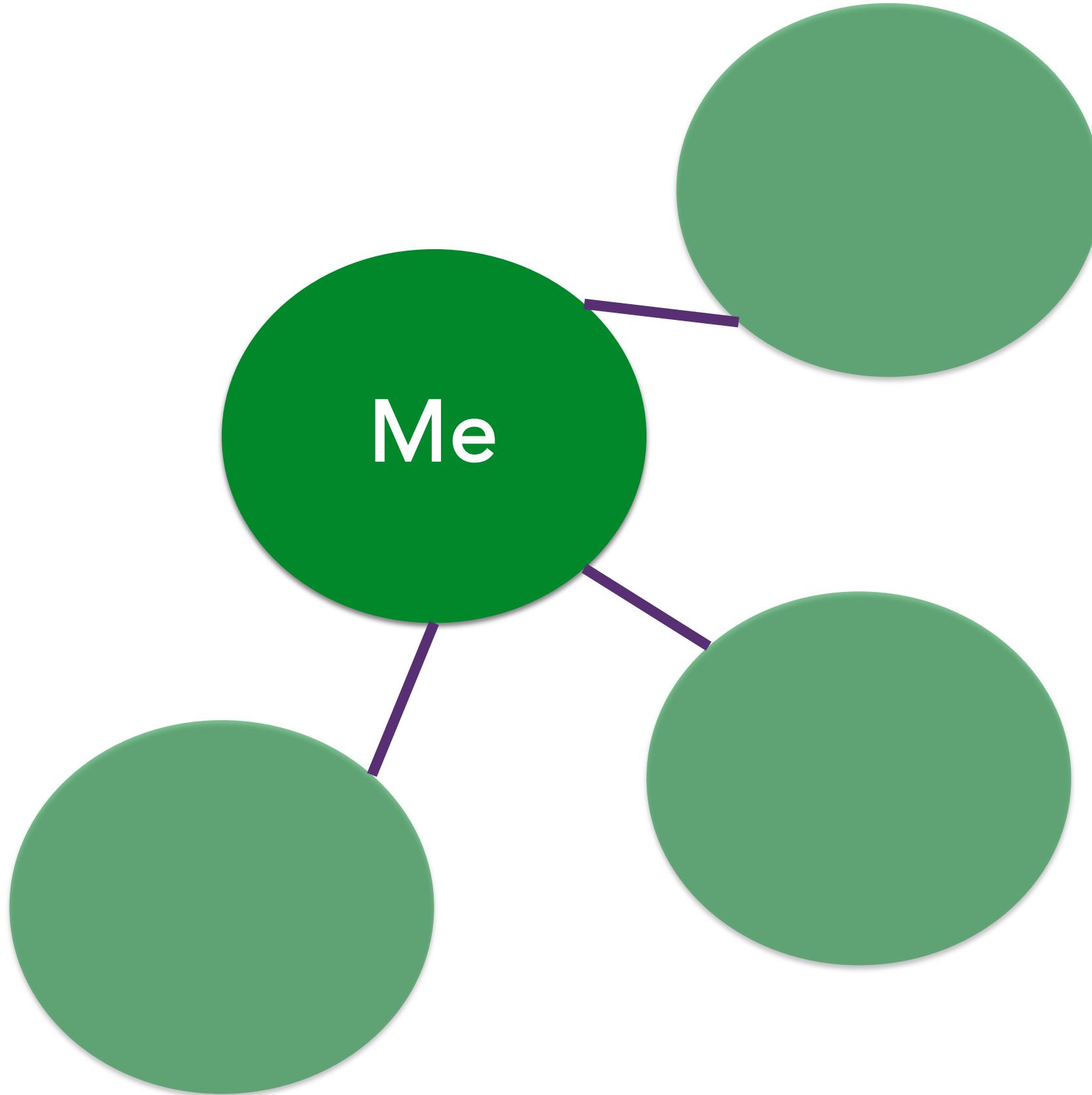


git

GitHub

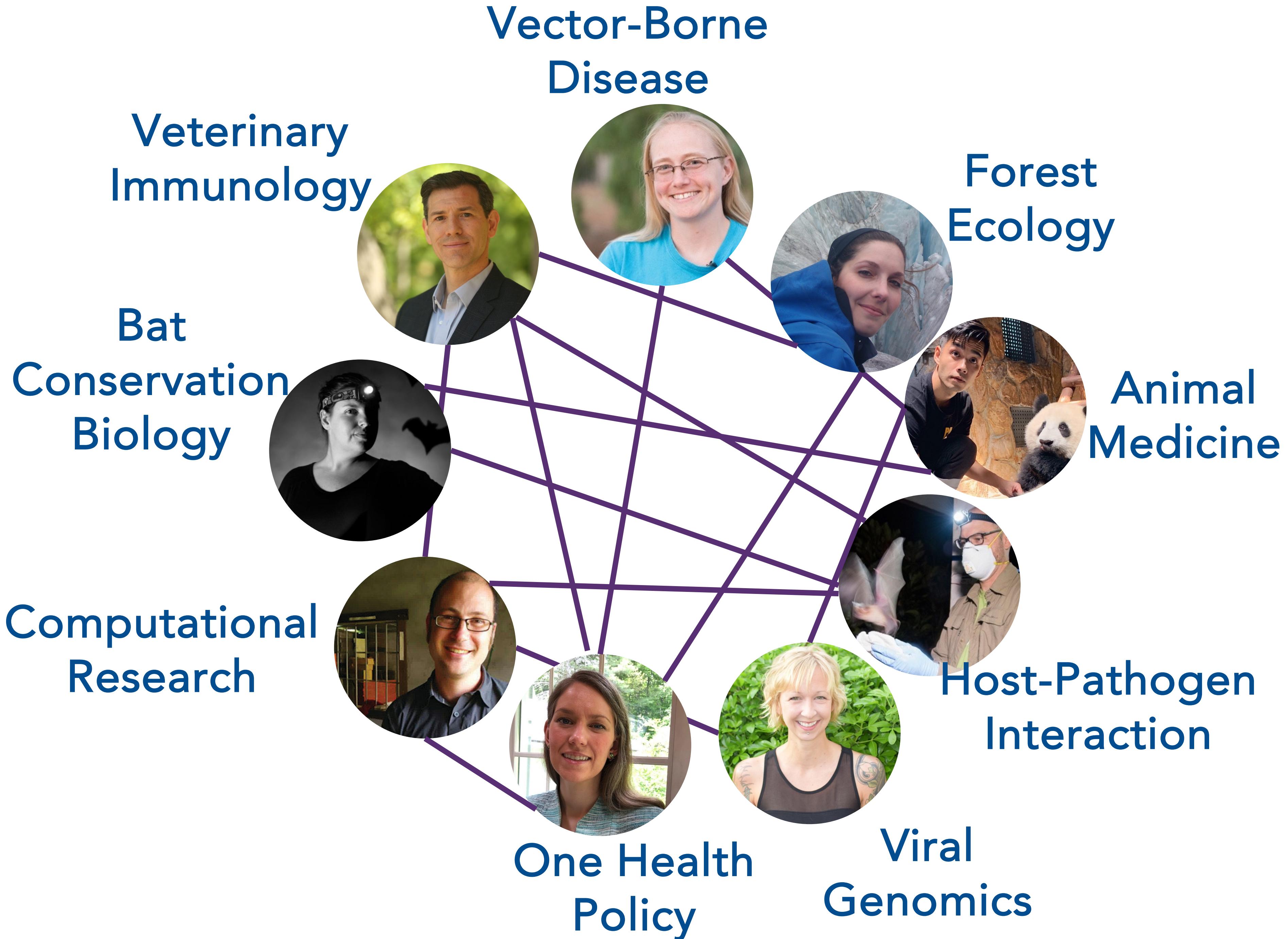


How to Scale these Practices and Tools?



What Problems Arise at Team or Organization Scale?

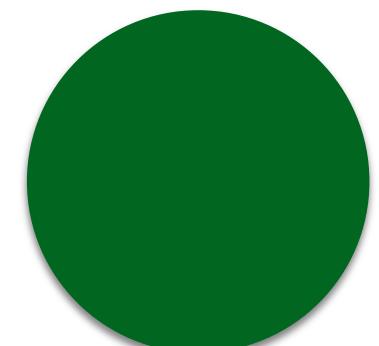
- How do we collaborate efficiently within and across disciplines?
- How do we share best practices and maintain common standards of quality?
- How do we re-use information and work across projects simultaneously and in time?
- How do we share work while protecting privacy and security?



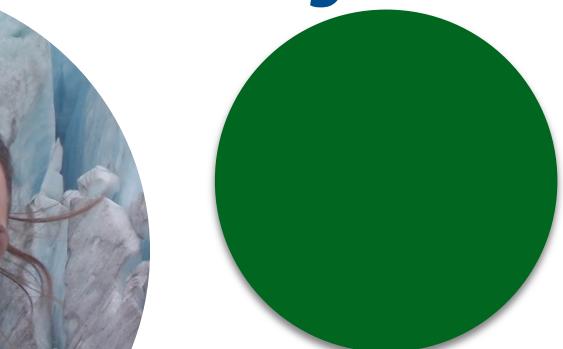
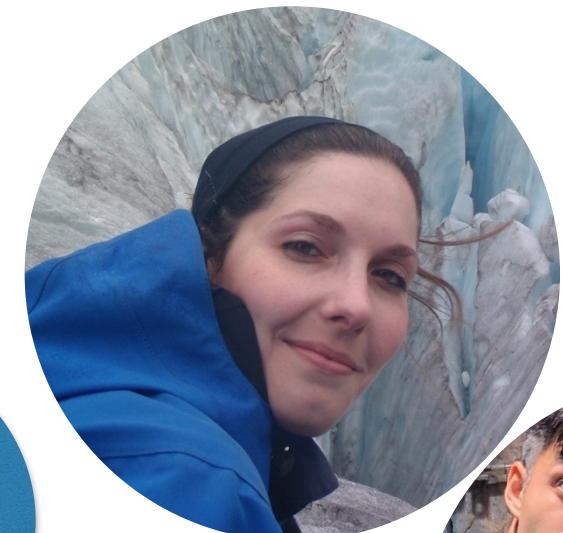
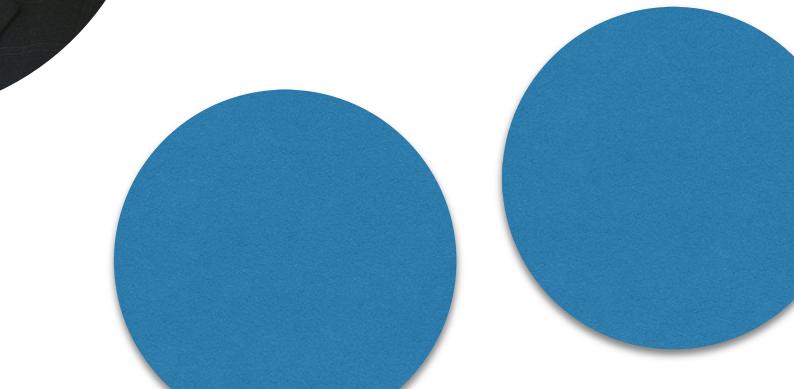
- An interdisciplinary group of collaborators working on common scientific, conservation, and public health issues
- 50-60 staff, most scientific

Epidemiological Simulation

Host Process Models



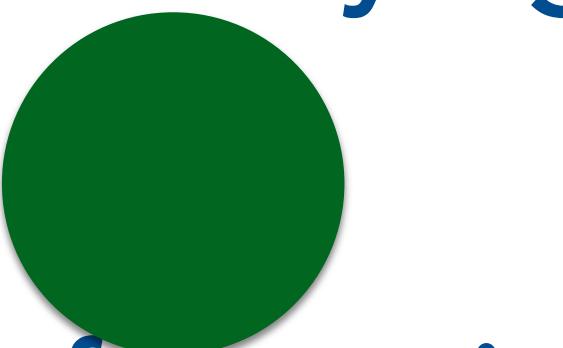
Community Analysis



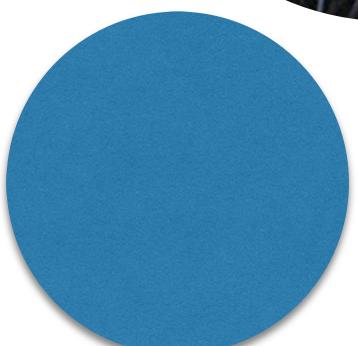
Spatial Analysis



Phylogenetics



Machine Learning and forecasting



Data Visualization

Bioinformatics

- 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

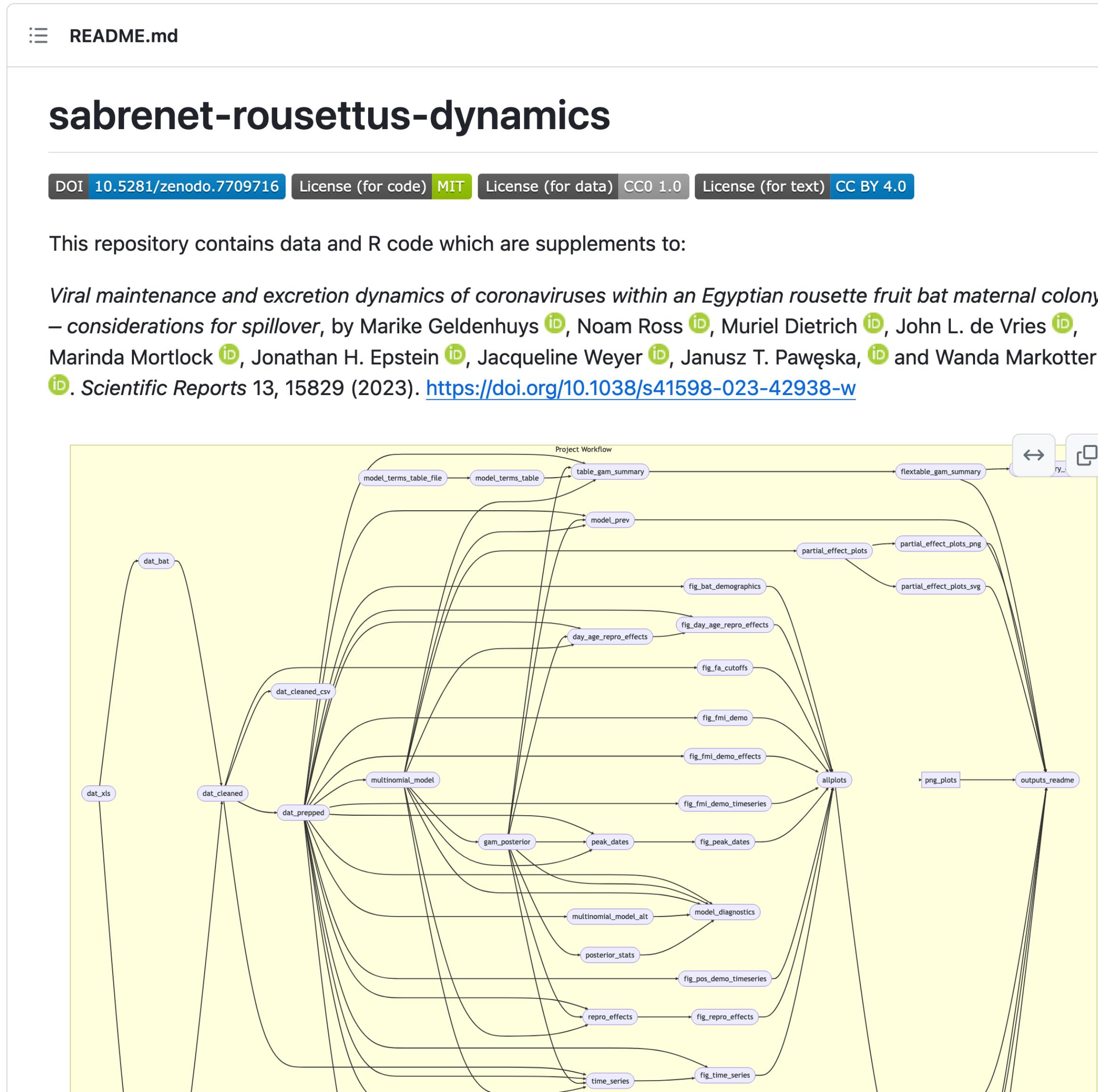
A Common Workflow Language

(Not THE “Common Workflow Language”)



- {targets} is an R-based build system for managing workflows and caching large computations
- Imposes an 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

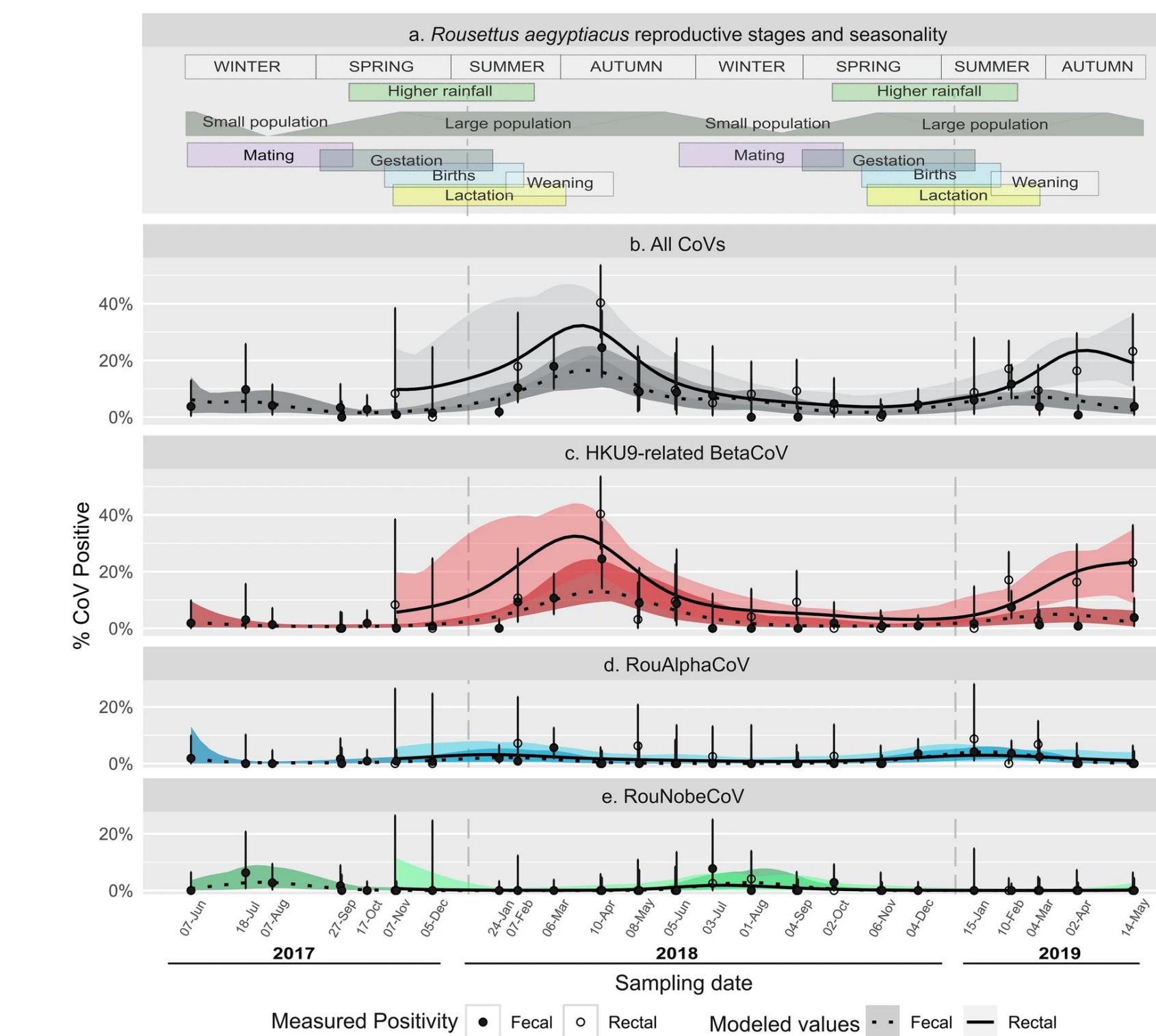


Article | [Open access](#) | Published: 22 September 2023

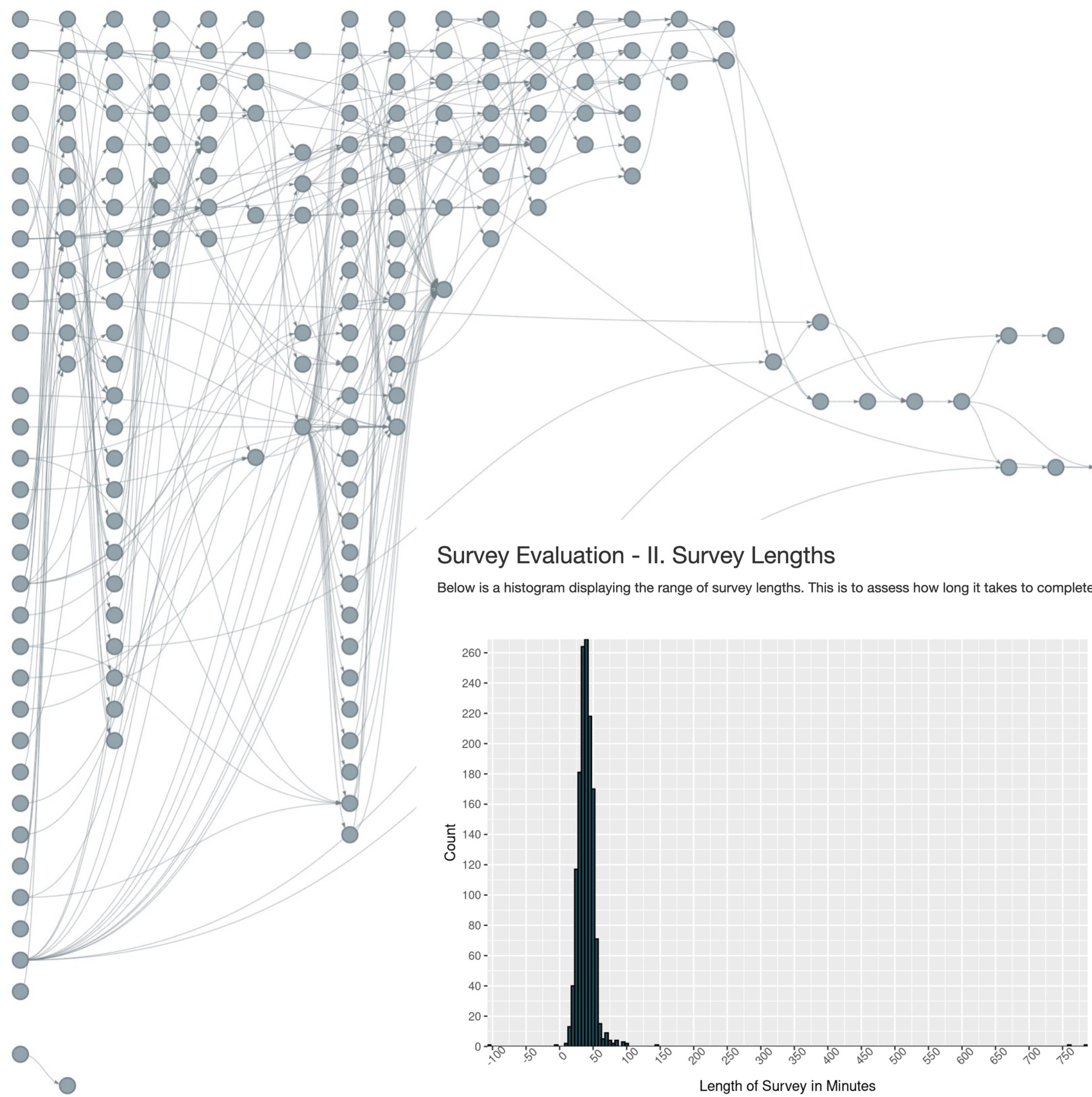
Viral maintenance and excretion dynamics of coronaviruses within an Egyptian rousette fruit bat maternal colony: considerations for spillover

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

Scientific Reports 13, Article number: 15829 (2023) | [Cite this article](#)



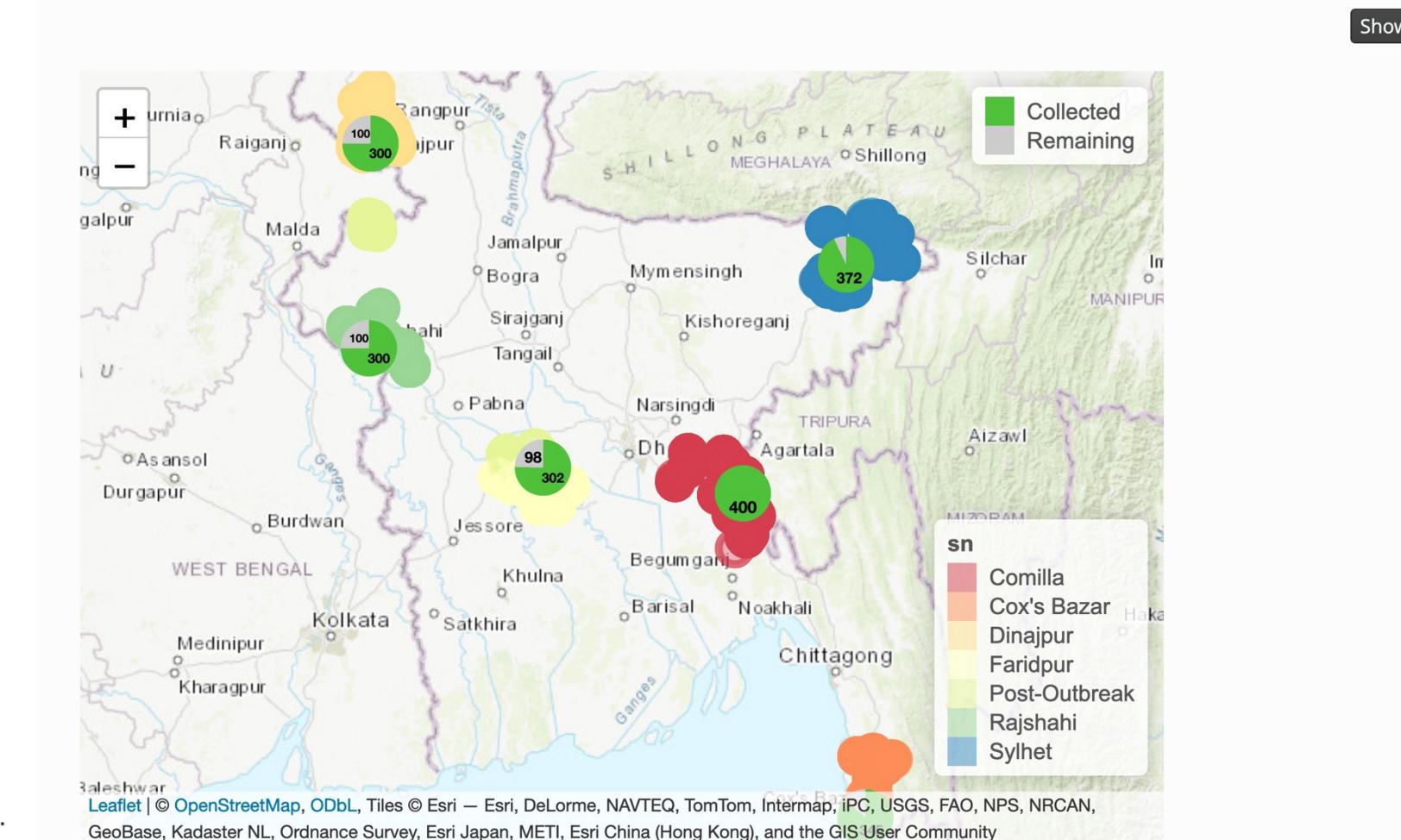
A Common Workflow Language



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.



Plates Recorded Issue Types Records Added

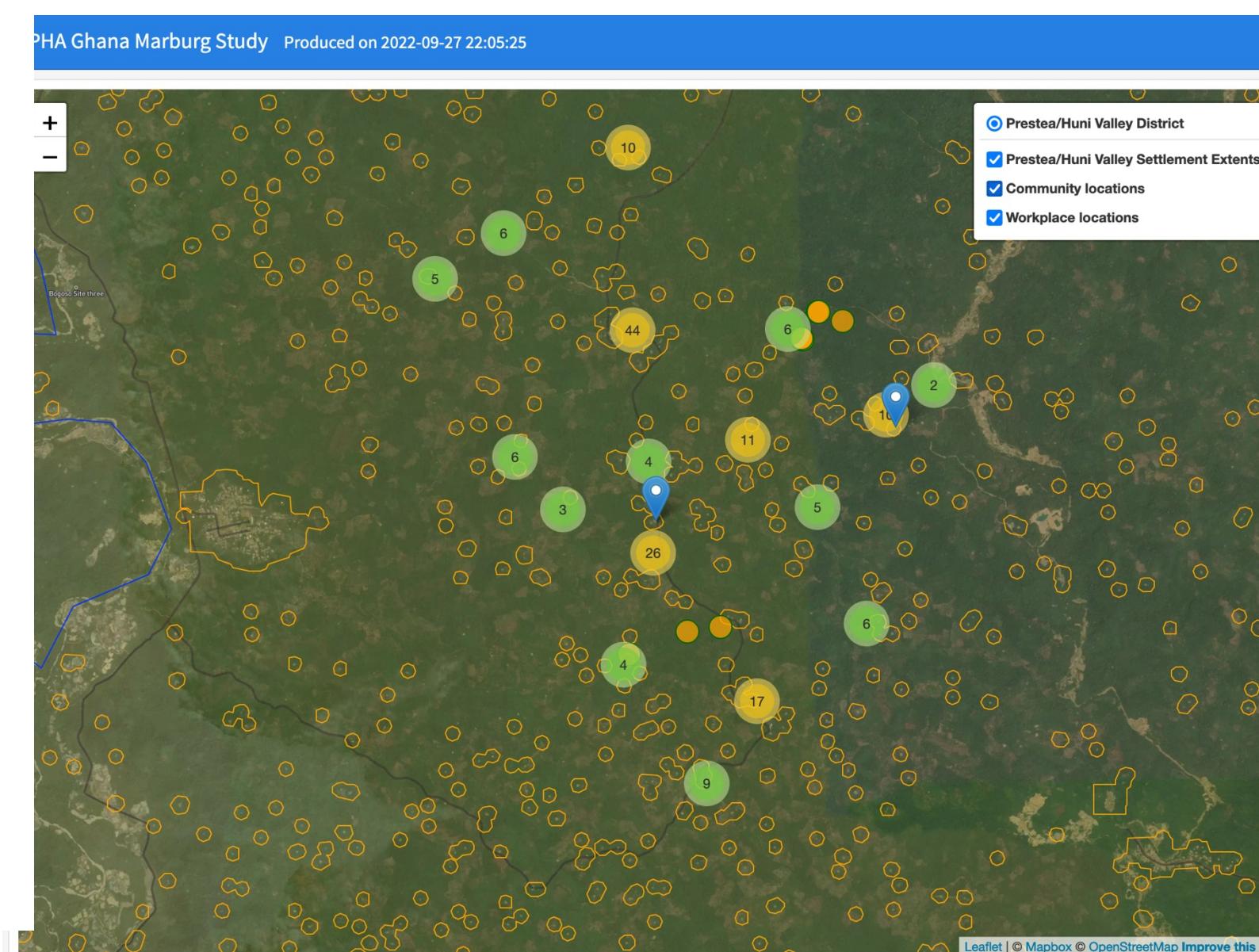
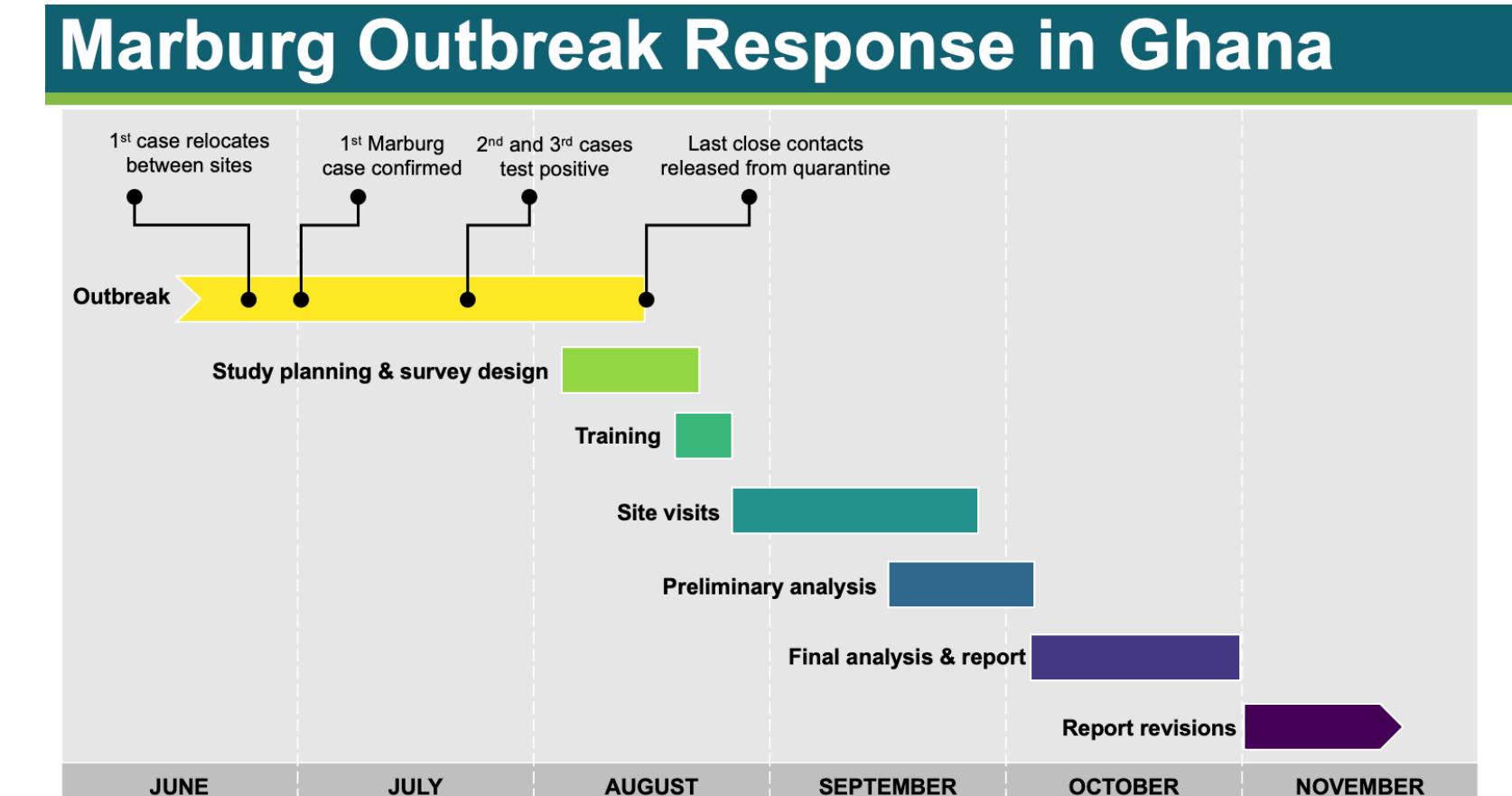
0 (0%)

7

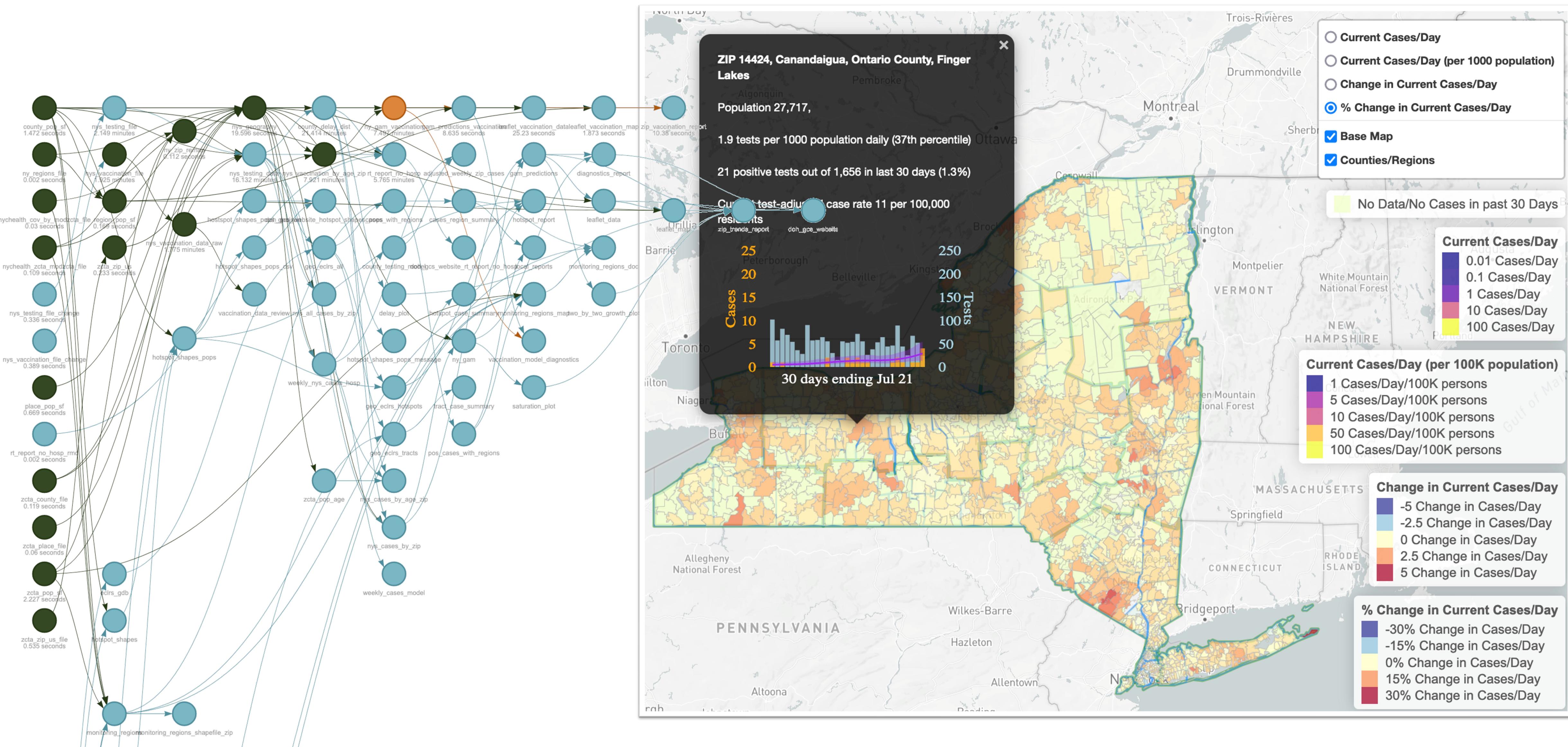
NA (NA)

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

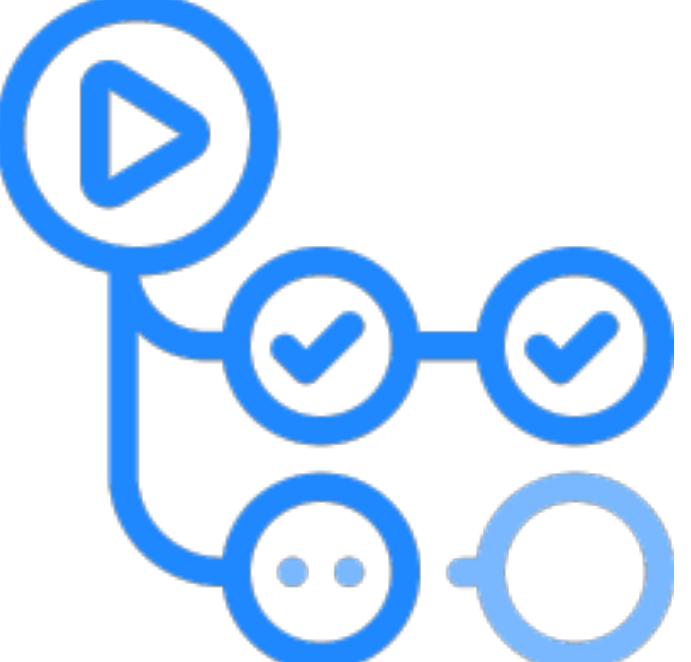
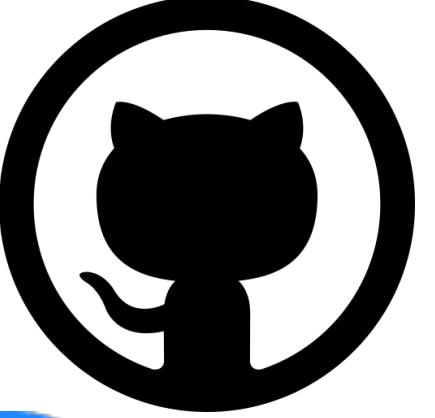
A Common Workflow Language



A Common Workflow Language



Process Automation – GitHub Actions



ecohealthalliance / nipah-bangladesh-automation

Type ⌘ to search

Code Issues 23 Pull requests Actions Projects Wiki Settings

← pcr-automations

✓ pcr-automations #75 Re-run all jobs ...

Summary

Jobs

✓ pcr-automations

Run details

Usage

Workflow file

pcr-automations succeeded 2 days ago in 4m 10s

Search logs

- > ✓ Set up job 13s
- > ✓ Pull ghcr.io/rtcamp/action-slack-notify:v2.2.0 4s
- > ✓ Initialize containers 55s
- > ✓ update permissions for container based workflows 0s
- > ✓ Run actions/checkout@v2 2s
- > ✓ Install system dependencies 21s
- > ✓ Decrypt repository using symmetric key 0s
- > ✓ Install packages from renv.lock (with cache) 42s
- ✓ Install packages from renv.lock (local, no cache) 0s
- > ✓ Run targets workflow to deploy, archive and email reports 1m 48s
- ✓ Run targets workflow without deploy and archive of reports (local with ACT) 0s
- > ✓ Show link to automation reports 0s
- > ✓ Slack Notification 0s

Managing Privacy/Security/Openness Trade-offs

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



DMPTool

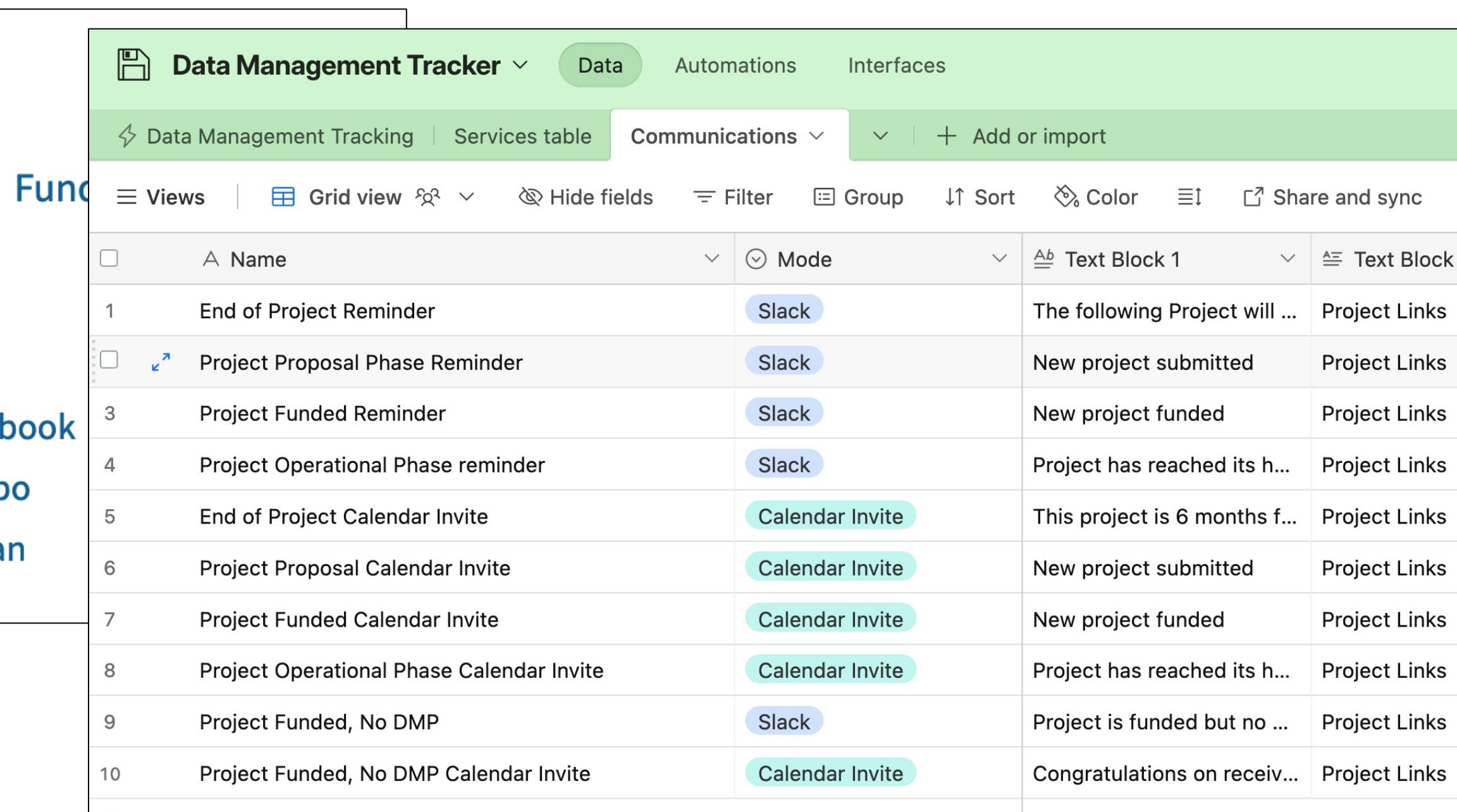
Build your Data Management Plan



EcoHealth Alliance

My Dashboard Create Plan Fund

- [EcoHealth Alliance](#)
- [EHA Github](#)
- [EHA Modeling & Analytics Handbook](#)
- [EHA Data Management Plan Repo](#)
- [EcoHealth Alliance Data Librarian](#)

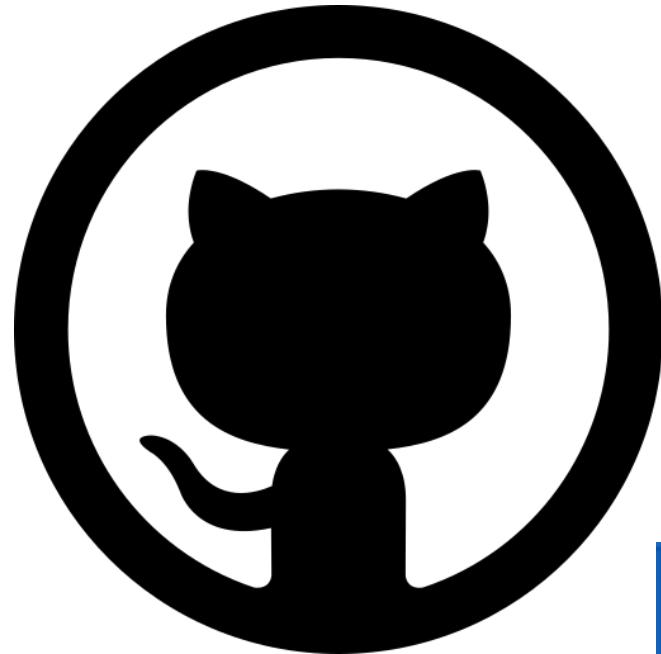


The screenshot shows a "Data Management Tracker" interface with a green header bar. The header includes the title "Data Management Tracker", tabs for "Data", "Automations", and "Interfaces", and buttons for "Add or import" and "Share and sync". Below the header is a toolbar with icons for "Views", "Grid view", "Hide fields", "Filter", "Group", "Sort", "Color", and "Share and sync". The main area is a grid table with 10 rows, each representing a communication task. The columns are: "A Name" (checkbox), "Mode" (dropdown), "Text Block 1" (dropdown), and "Text Block" (dropdown). The rows contain the following data:

A Name	Mode	Text Block 1	Text Block
1 End of Project Reminder	Slack	The following Project will ...	Project Links
2 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

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



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 \(agwa@andrewayer.name\)](mailto:Andrew Ayer (agwa@andrewayer.name)). For more information, see <https://www.agwa.name/projects/git-crypt>.

- 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



- Tablet-based field data collection platform
- Designed for robustness and offline environments
- Open source, self-run on cloud servers (service available)
- 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)	
Follow_up_questionnaire >	Follow_up_questionnaire 20230405	2,209 Submissions > (last Monday 04:34)	
Follow_up_questionnaire_draft >	Follow_up_questionnaire_draft	1 Submission > (last 2022/04/26 14:34)	
HQ_Test >	HQ_Test 20220207	10 Submissions > (last 2022/02/07 10:17)	
Human_Questionnaire >	Human_Questionnaire 20220207	499 Submissions > (last 2022/10/01 02:42)	
Site_Interface_Characterization_Form >	Site_Interface_Characterization_Form 20220511	276 Submissions > (last 2023/10/25 11:31)	
Workshop Demonstration of ODK >	workshop_survey 20220207	31 Submissions > (last 2022/10/18 04:31)	

Data Platforms for all Disciplines

ANIMAL CARE & USE PROGRAM		Data	Automations	Interfaces		
		ORDER	FAMILY	GENUS ETC	IACUC TEXT	GRANT TEXT
ORDER	FAMILY	GENUS ETC	IACUC TEXT	GRANT TEXT		
5 Net, Handheld	In Progress	Carnivora Galliformes (Fowl) Pholidota (Pangolins) Primates & ALL SPECIES: OUTBREAK	Manidae Cercopithecidae Hominidae Lorisidae Hylobatidae Tarsiidae Pteropodidae Felidae	Asellia Miniopterus Scotozous Chaerephon Pipistrellus Megaderma Ptilocercus Scotophilus Vespertilio Pteropus	BATS: we may occasionally extract bats from roosts with hand-held hoop ("butterfly") nets upon exit or during flight. This method will ...	
6 Net, Mist, Standard	In Progress	& ALL SPECIES: OUTBREAK	Pteropodidae Chiroptera (Bats)	Asellia Miniopterus Nycteridae Molossidae Miniopteridae Hipposideridae Megadermatidae	Scotozous Chaerephon Pipistrellus Megaderma Scotophilus Vespertilio Pteropus Plecotus	BATS: will be captured using either harp traps or mist nets set in flyways or at the entrance to caves or other structures in which bats are roosting....
7 Net, Long	In Progress	& ALL SPECIES: OUTBREAK	Leporidae Lagomorpha (Hares & Rabbits)			HARES: we will use nets that are 90-100 m in length and approximately 1 m high are set, then the hares are driven into the net by people ...
8 Net, Mist, Triple-High	In Progress	& ALL SPECIES: OUTBREAK	Pteropodidae Chiroptera (Bats)	Asellia Miniopterus Nycteridae Molossidae Miniopteridae Hipposideridae Megadermatidae	Scotozous Chaerephon Pipistrellus Megaderma Scotophilus Vespertilio Pteropus Plecotus	BATS: will be captured using a large canopy net (30' x 30') or triple high mist net system that is suspended by rope between two aluminum ...
9 Trap, Harp	In Progress	& ALL SPECIES: OUTBREAK	Pteropodidae Chiroptera (Bats)	Asellia Miniopterus Nycteridae Molossidae Miniopteridae Hipposideridae Megadermatidae	Scotozous Chaerephon Pipistrellus Megaderma Scotophilus Vespertilio Pteropus Plecotus	BATS: will be captured using either harp traps or mist nets set in flyways or at the entrance to caves or other structures in which bats are roosting....
10 Trap, Pitfall	Not Started	Didelphimorphia (Opossums) Macroscelidia (Elephant Shrews) & ALL SPECIES: OUTBREAK Eulipotyphla (Hedgehogs, Solenodon) Rodentia	Didelphidae Nesomyidae Pedetidae Muridae Soricidae Sciuridae Erinaceidae Talpidae Bathyergidae Cricetidae	Bathyergus Mesocricetus Didelphis Glis		
11 Trap, Sherman	Complete	Scadentia (Treeshrews) Afrosoricida (Moles, Otter-shrews) Macroscelidia (Elephant Shrews) Didelphimorphia (Opossums) & ALL SPECIES: OUTBREAK	Potamogalidae Didelphidae Nesomyidae Ptilocercidae Pedetidae Muridae Soricidae Sciuridae Erinaceidae	Ptilocercus Bathyergus Mesocricetus Didelphis Glis	IACUC TEXT: Free-ranging rodents, moles, and shrews will be capture traps (Sherm)	GRANT TEXT: Just before dusk, Sherman traps will be set and baited with
12 Trap, Tomahawk	Complete	Carnivora Afrosoricida (Moles, Otter-shrews) Macroscelidia (Elephant Shrews) Didelphimorphia (Opossums) Scadentia (Treeshrews)	Potamogalidae Felidae Didelphidae Nesomyidae Ptilocercidae Pedetidae Muridae Procyonidae Soricidae Mustelidae	Ptilocercus Bathyergus Procyonis Mesocricetus Didelphis Glis Prionodon	IACUC ranging and shr capture traps (Sherm, Tomhawk ...)	from both humans and ...



ecohealthalliance / airtabler

Process Automation

EHA Science Talks Data Automations Interfaces

Automations List ON Schedule M3 Meetings on Cal... ⓘ Run History Test Automation

Schedule M3 Meetings on C... Create calendar events for M3 me...
Reminder to send M3 invites When a record matches conditions, ...
Automation 1 No description OFF

TRIGGER When a record is updated Name, Date / Time, Type, and 1 more field

ACTIONS

If Type has any of Methods and Models , autocal , Date / Time is on or after today, End Time is not empty, Name is not empty, and Calendar ID is empty
Make calendar events for new M3 events

Google Calendar: Create event

Update record

Otherwise if Type has any of Methods and Models , autocal , Calendar ID is not empty, Date / Time is on or after today, and End Time is not empty
Update a Google Calendar event

Google Calendar: Update event

+ Add condition

✓ Review test results

The screenshot shows the Airtable interface for process automation. At the top, there's a navigation bar with tabs for 'EHA Science Talks', 'Data', 'Automations' (which is selected), and 'Interfaces'. Below the navigation is a toolbar with buttons for 'Automations List', 'Run History', and 'Test Automation'. The main area displays a list of automations. One automation is currently selected: 'Schedule M3 Meetings on Cal...', which is set to 'ON'. This automation triggers when a record is updated, specifically for 'Name, Date / Time, Type, and 1 more field'. The trigger is shown with a green checkmark icon. The automation then branches into two paths based on the 'Type' field. The first path handles cases where 'Type' includes 'Methods and Models', 'autocal', and either 'Date / Time is on or after today' or 'End Time is not empty'. It involves creating calendar events for new M3 events using Google Calendar and updating the record. The second path handles cases where 'Type' includes 'Methods and Models', 'autocal', and both 'Calendar ID is not empty' and 'Date / Time is on or after today'. It involves updating a Google Calendar event. Both paths lead to a final step: 'Review test results', indicated by a checkmark icon.

Cross-Training Across Teams



Introduction to modern scholarly works databases for literature reviews and more
File | /Users/collinschwantes/Documents/research-output-catalog/presentations/m3_oa_semscholar.html#13

OpenAlex Use Case: Find papers about Rift Valley Fever

Concepts - wikidata categories that are applied to works based on title and abstract

```
 cwd_concept <- openalexR::oa_fetch(entity = "concepts", search = "Rift Valley Fever")
 cwd_concept$works_api_url #query for all works

## [1] "https://api.openalex.org/works?filter=concepts.id:C2778960357"

 cwd_concept %>%
   select(display_name, description, works_count)

## # A tibble: 1 × 3
##   display_name      description  works_count
##   <chr>            <chr>          <int>
## 1 Rift Valley fever human disease      3439
```

 EcoHealth Alliance

13 / 16

- 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

The screenshot shows a screenshot of an Airtable database titled "EHA Science Talks". The database has a green header bar with tabs for "Data", "Automations", and "Interfaces". Below the header, there are navigation links for "Talks", "Speakers", "Meta Data", "Description", and a "+" button. The main view displays a table with columns: "Name", "Speaker", "Keywords", "Abstract", and "Recording". The table contains 7 rows of data, each representing a recorded talk. The "Keywords" column uses color-coded tags to categorize the topics. A sidebar on the left lists various categories such as "Scheduled EHA Talks", "Proposed Topics and Guests", "Recorded Talks", "Calendar", "External Talks", "Intentionally Blank", "Methods and Models Meetings", "Airtable, Data Management, an...", "Form to suggest speakers for E...", "ODK Talks", "Reproducibility best practices" (which is selected), "Targets Talks", "Airtable Talks", "Submit Talk Materials", and "Personal views".

Name	Speaker	Keywords	Abstract	Recording
M3: Airtable Backups	Collin Schwantes	airtable data management	This week at M3 Collin will be walking through backing up project AirTable databases locally and to ...	https://drive.google.com/file/d/1lIQX3x390i5qhnYKangmOOQgVpGGIBf1...
M3: Strategies for Project Close-Out	Collin Schwantes	reproducibility coding best practices data management	This week at M3 we will discuss a draft end-of-project strategy. Ending projects properly ensures...	https://drive.google.com/file/d/1IVZQOA03N1C9PBkn1Y6uxAE_Csm8...
Elements of Reproducibility- managing R packages with renv	Collin Schwantes	renv reproducibility coding best practices	In the fourth talk of Elements of Reproducibility series, Collin will give a review/tutorial of using th...	https://drive.google.com/file/d/1lhbkU8FxqZZ76URo3H9NHZJ-o0PV9PzX...
Elements of Reproducibility- Targets and reproducible workflows	Collin Schwantes	reproducibility targets coding best practices	In the third talk of the Elements of Reproducibility series, Collin will give a review/tutorial of using th...	https://drive.google.com/file/d/1lwbjwlwts9GuiticsAKinwVOCAWJBv59...
Elements of Reproducibility II- Git and GitHub	Collin Schwantes	git reproducibility version control coding best practices	In this second in our Elements of Reproducibility series, Collin will give a review/tutorial of using Gi...	https://drive.google.com/file/d/1m307XTB0dazFZzVogm59kolhff50kDl...
M3: Building Blocks of Reproducibility Part I	Collin Schwantes	coding best practices reproducibility	This week, Collin will kick off a series of M3's on "The Building Blocks of Reproducibility." Our goa...	https://drive.google.com/file/d/1m8yyYWFlaNnOZQ0Oi4xeBiM36KA_xmfs...
M3: Data Management Plans for All	Collin Schwantes	data management	At this M3, Collin will speak	https://drive.google.com

- Recorded talks live in a searchable catalog of topics

Cross-Training Across Teams

Monday, January 23rd ▾

12:07 PM [REDACTED] Hey everyone, I have a coding question. I know I can write a function to do this, but I feel like this should be common enough that there's probably a package or function that makes this easy, so I thought I'd ask:)
I have a dataframe that I need to do some cleaning on and then proceed with analysis. For the cleaning I'd like to convert all columns to character, then after the cleaning, I'd like to convert all columns back to their original class. Does anyone know a simple code to do this? Thanks!

12:39 PM [REDACTED] You could do this partially with `readr::type_convert()`
Here's an example.

```
library(tidyverse)
dat <- tibble(a = 1:2, b = c(TRUE, FALSE), c = factor(c(1,2)))
col_types <- paste(map_chr(dat, ~str_sub(class(.), 1, 1)), collapse = "") # get first initials of col types
dat_as_chr <- dat |> mutate(across(everything(), as.character))
type_convert(dat_as_chr, col_types)
```

12:41 PM **Nathan Layman** Dang beat me too it. I was going to suggest merge

```
test <- iris |> mutate_all(as.character) # change all columns to character class
test <- test |> slice(1:10) # remove some rows to approximate cleaning
test |> str()
merge(iris, test) |> str() # use merge to re-type test to iris's original column classes
```

1 thumbs up, 1 reply

5 replies Last reply 10 months ago

- 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



Modeling & Analytics Handbook

Introduction

1 Contributing

2 Quickstart for computing

3 Project Management

4 R and Reproducible Analysis

5 EHA Team Communication

6 Documentation and Outputs

7 Data Management

8 Airtable

9 Version Control, Git and Github

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 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

Bioinfo training #99

Merged by collinschwantes ecohealthalliance:master ← alexarmerov:master on Jun 5 v1.0.0

Conversation 0 Commits 1 Checks 0 Files changed 1

all commits File filter Conversations Jump to 0 / 1 files viewed

10 training-resources-and-plans.Rmd

83 83 @@ -83,4 +83,14 @@ some of the material from Geomputation in R in less depth, and gives you fewer t
- [Leaflet for R](<https://rstudio.github.io/leaflet/>) is a manual on the use of the R **leaflet** package to harness Le
open-source JS library for creating interactive maps. Leaflet maps particularly useful for exploring and visualizing sp
into R Markdown documents. You should take a course or have knowledge of R Markdown prior to taking this course.

84 84

85 85

86 **### Bioinformatics**

87 - Conceptual and practical introduction to some of the main topics in Bioinformatics.

88

89 **## Metagenomics**

90 - [Quality control](https://github.com/alexarmerov/Metagenomics/blob/main/Docs/Quality_control.md)

91 - [Assembly](<https://github.com/alexarmerov/Metagenomics/blob/main/Docs/Assembly.md>)

92 - [Alignment](<https://github.com/alexarmerov/Metagenomics/blob/main/Docs/Alignment.md>)

93

94

95

96

- 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

create predict API functions #152

Merged dev feature/prediction-api 15 hours ago

Conversation 1 Commits 4 Checks 0 Files changed 2

ernestguevarra commented yesterday
addresses #74

create predict API functions Verified 5d84345

ernestguevarra added the feature engineering label yesterday

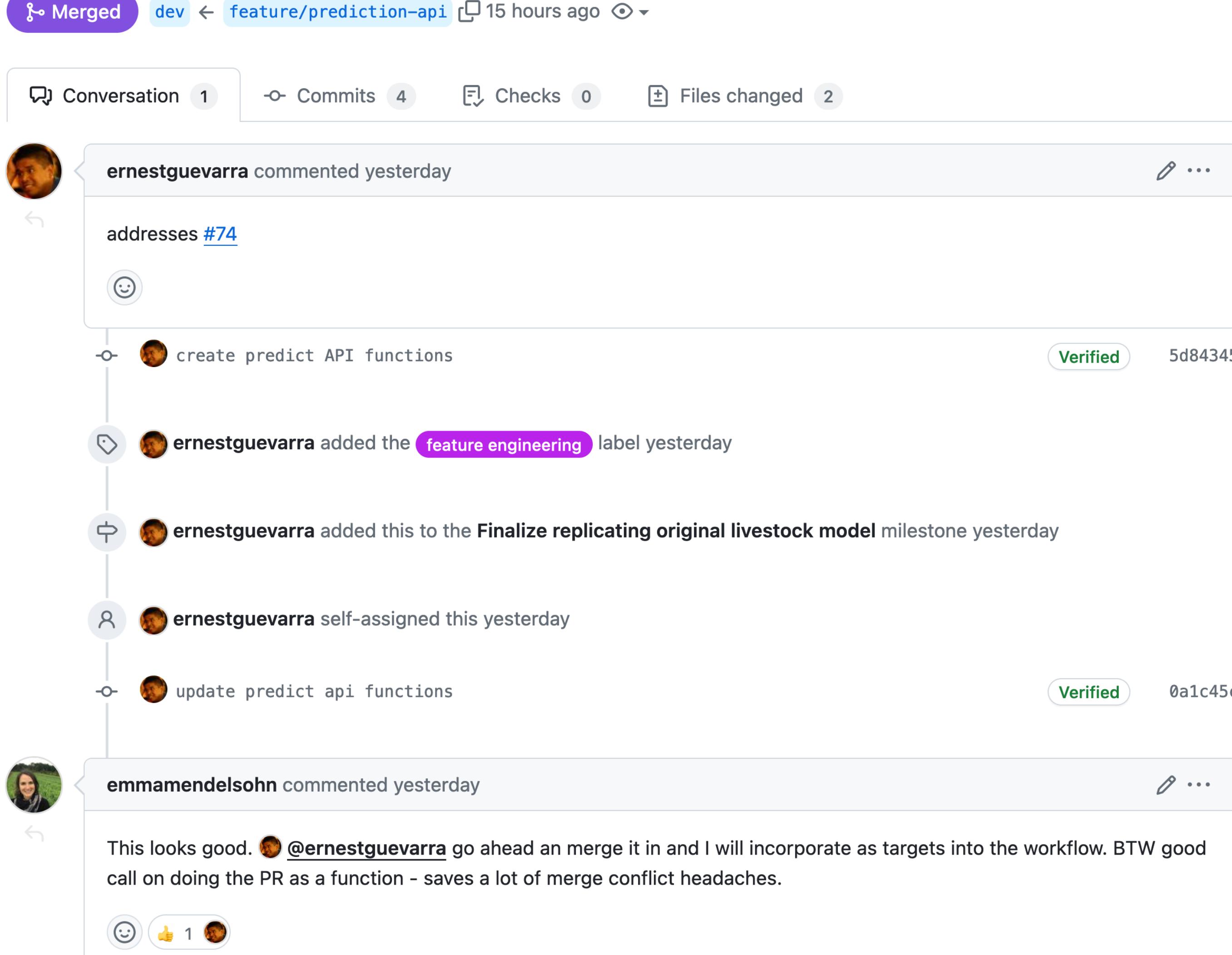
ernestguevarra added this to the Finalize replicating original livestock model milestone yesterday

ernestguevarra self-assigned this yesterday

update predict api functions Verified 0a1c45c

emmamendelsohn commented yesterday
This looks good. @ernestguevarra go ahead and merge it in and I will incorporate as targets into the workflow. BTW good call on doing the PR as a function - saves a lot of merge conflict headaches.

1



- 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.org