# Facilitating programmatic data retrieval from CAP with R and Python clients: rCAX and pycax packages

## a step on a journey towards Transforming to Open Science workflows

**Eli Holmes,** NMFS Open Science, Lead

NOAA Fisheries
Northwest Fisheries Science Center
Math Bio Program


jointly developed with Mari Williams

*https://nmfs-opensci.github.io/*

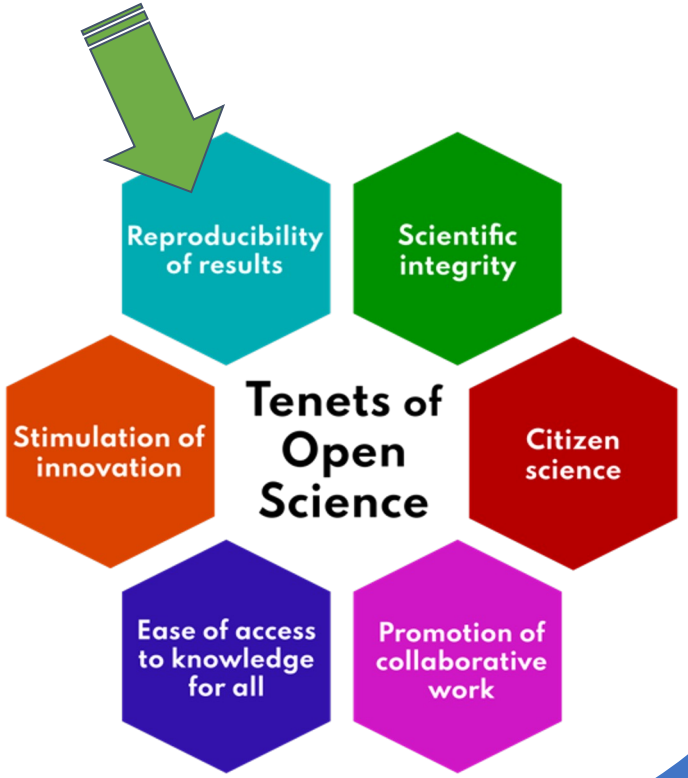
NMFS OPEN SCIENCE



NOAA FISHERIES

## Reproducibility is hard
## Reproducibility is not a given

*Recently scientific studies have shown that significant (over half) of studies cannot be replicated – even with the raw data and a narrative of the methods.*




# Reproducibility is Core Objective of Open Science



NOAA FISHERIES

## Data

- Poor documentation
- Partial methods
- Unknown provenance
- No metadata
- No contact information
- No context

## Analyses

- Analyses, plots, tables with no documentation
- Manual undocumented manipulations
- Many data file in different formats
- Scripts of various analyses
- Emails, emails, emails
- Lots of Google docs
- Files on individual folders
- Each team member working solo


## Semi-reproducible product:

- Paper
- Decision
- Report

Very time-consuming to reproduce when we have updated data or discover errors



NOAA FISHERIES

# As time goes on, reproducibility decreases as both the data and analysis "degrade"

Emails get lost and deleted

The manual steps are forgotten slowly but surely.

The people who did the analysis move on or retire

The biologists who know the original data move on or retire

**Unreproducible result (with no hope to replicate or fix)**
**Lost data (no longer able to understand)**

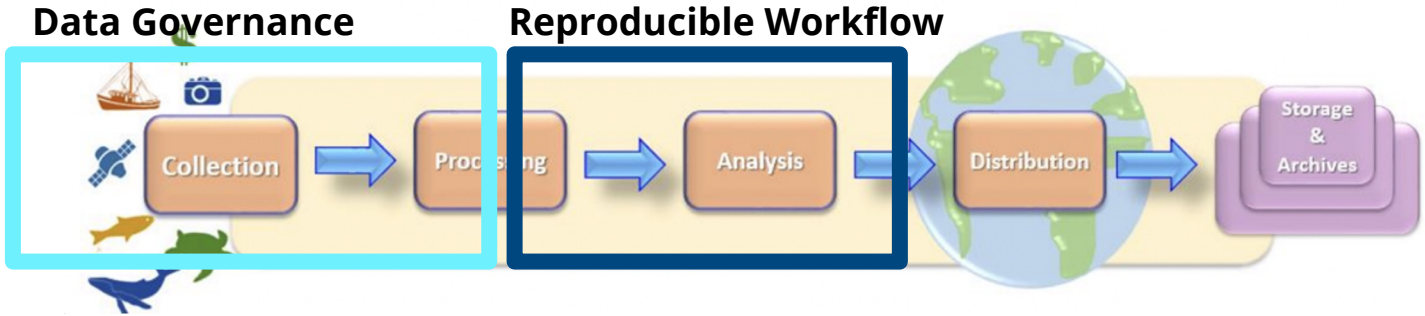


NOAA FISHERIES

# How does one create a “reproducible scientific pipeline”?

- **Data Governance**: Data management and documentation
- **Data wrangling**: Eliminating manual manipulation of data
- **Analysis:** A documented pipeline rather than a patchwork of poorly documented analyses.

- **Scripts** instead of manual steps
- **Version-control**: all changes and decision documented
- **Text and code integrated**
- A “make” file that reproduces the product
- A “devcontainer” of the environment




[Fisheries Information Management Modernization Workshop 2020, Tech Memo](#) September 17-19, 2019, NMFS Office of Science and Technology (OST)


Gentemann, Chelle L., Shrestha, Sudhir, Ivey, Yvonne, & Hall, Cynthia. (2023, February 9). TOPS February 9 Community Forum. Zenodo. https://doi.org/10.5281/zenodo.7626005

# NMFS Openscapes training in Open Science

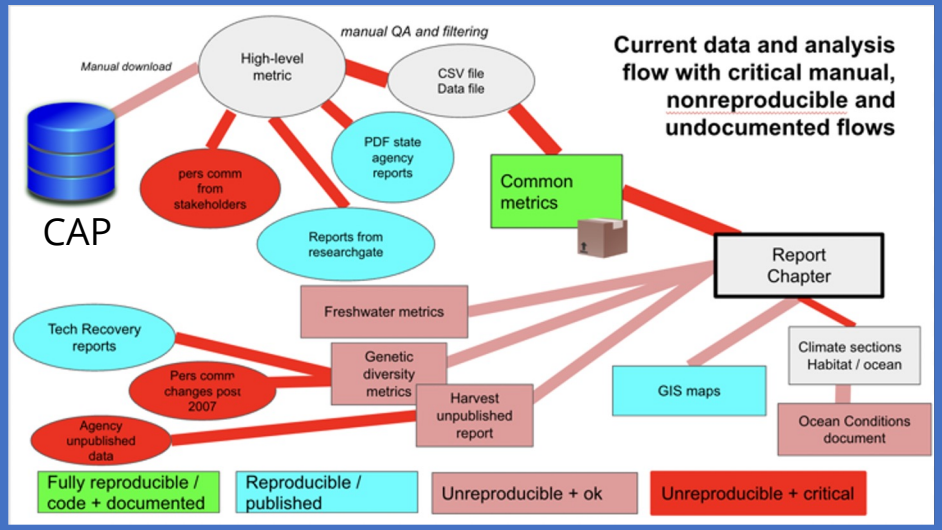

At NMFS, a grassroots effort staff training in Open Science

9 NMFS Champions Cohorts (ca 300 staff)

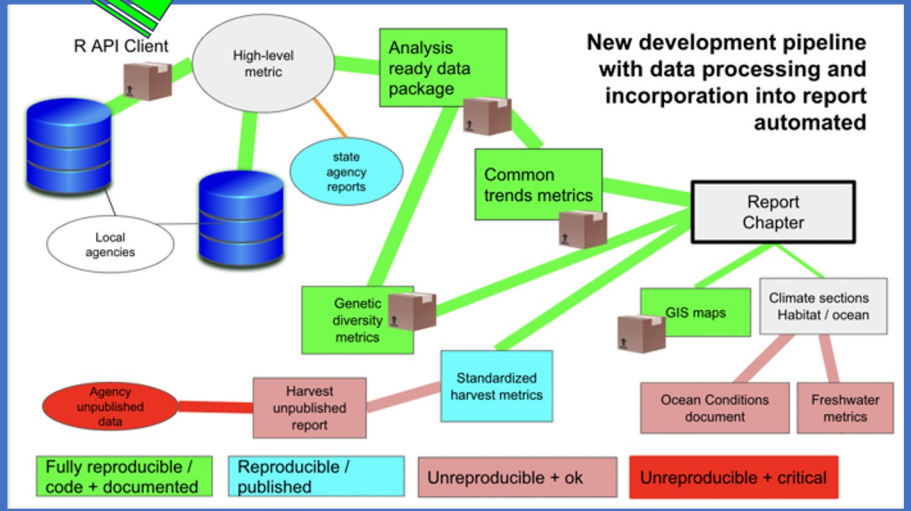
- 2020: Winter NEFSC
- 2021: Spring NWFSC
- 2021: Fall NWFSC, AFSC, SEFSC, NEFSC
- 2022: Winter AFSC
- 2022: Summer SEFSC/SERO
- 2022 Fall 4 cohorts 6 science ctrs, WCRO

https://nmfs-openscapes.github.io/

# PNW Salmonid Viability Report (NWFSC) + Status Reviews (WCRO) Team




TASK


Data “Pathway” or “Journey” work by teams – with trained data pathway facilitators

- Team psychological safety work
- 3-4 co-work sessions
- Identify doable steps
- START MOVING FORWARD

# Programmatic data retrieval from CAP with R and Python clients: rCAX and pycax packages

Holmes EE, Williams M (2023). “rCAX: Coordinated Assessments REST API R Client. R package version 1.0.1.” NOAA Fisheries, Northwest Fisheries Science Center, Math Bio Program. **doi:10.5281/zenodo.7402463**.

https://nwfsc-math-bio.github.io/rCAX/

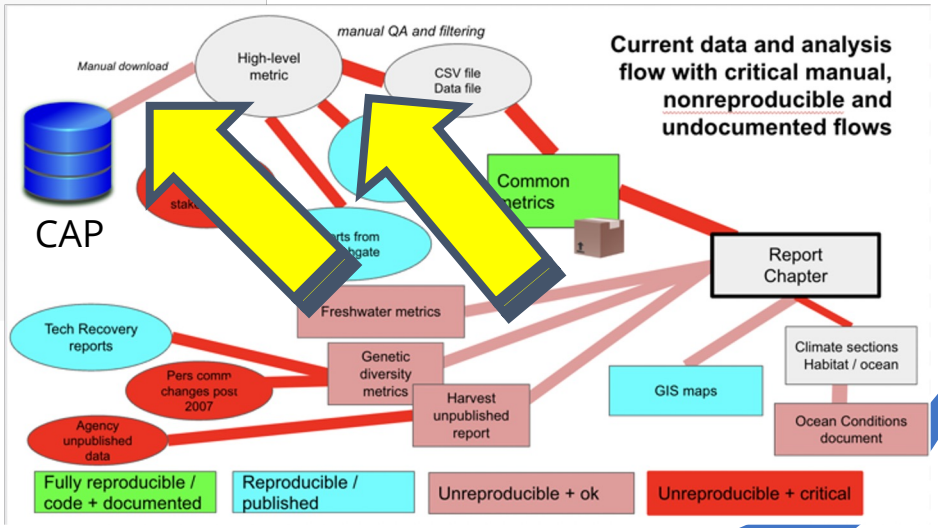



NOAA FISHERIES

# Why is programmatic access key for our workflows?

```
tab <- rcax_hli("NOSA",
  flist = list(nmfs_popid = 7),
  cols=c("nmfs_popid", "spawningyear", "tsaej", "nosaej"))
head(tab)
#>   nmfs_popid spawningyear tsaej nosaej
#> 1          7         1964  3020
#> 2          7         1965  2539
#> 3          7         1966  2984
#> 4          7         1967  3182
#> 5          7         1968  3594
#> 6          7         1969  2973
```

Automate the Q&A
**Make sure metadata retained**





**An open API with documentation!**



NOAA FISHERIES

Contribute!

# rCAX


`rCAX` is an R client for the **Coordinated Assessments API**. Coordinated Assessments data eXchange (CAX) is developed by the Coordinated Assessments Partnership (CAP). CAP is a collaborative process to efficiently share and provide access to standardized derived information, such as fish population-scale high-level indicators (HLIs) and supporting metrics. Participants in CAP include state fish and wildlife management agencies, tribes, federal agencies such as National Oceanic and Atmospheric Administration Fisheries (NOAA Fisheries) and Bonneville Power Administration (BPA), and others. CAP is co-sponsored by StreamNet and Pacific Northwest Aquatic Monitoring Partnership (PNAMP). Make sure to review the **StreamNet Terms of Use** for these data, the **StreamNet Data Policy** and the citation information from **StreamNet** and **PNAMP** for database queries.

## Installation

Install the latest GitHub release. You only need to do this once.

```
install.packages("remotes") # needed for the next line
remotes:::install_github("nwfsc-math-bio/rCAX@*release")
```

## Download a table

Read about the **CAX data and terms**. Then read the **Basic functions vignette** to get started and see examples of how to make queries.

To retrieve the NOSA data for NMFS_popid 7 that is the same as the file that one can download from **CAP Fish HLIs Tabular Query**, use

### Links

**Browse source code**

**Report a bug**

### License

**MIT** + file **LICENSE**

### Citation

**Citing rCAX**

### Developers

**Eli Holmes**
Author, maintainer

**Mari Williams**
Author

**Katie Barnas**
Contributor

### Dev status

GitHub v1.0.1

R-CMD-check passing

# pycax

Build and Deploy docs passing | Tests passing

pycax is an Python client for the Coordinated Assessments API. Make sure to review the StreamNet Terms of Use for these data, the StreamNet Data Policy and the citation information for database queries. pycax was developed by the Northwest Fisheries Science Center Math Bio Program.

NWFSC Math Bio CAX REST API clients:

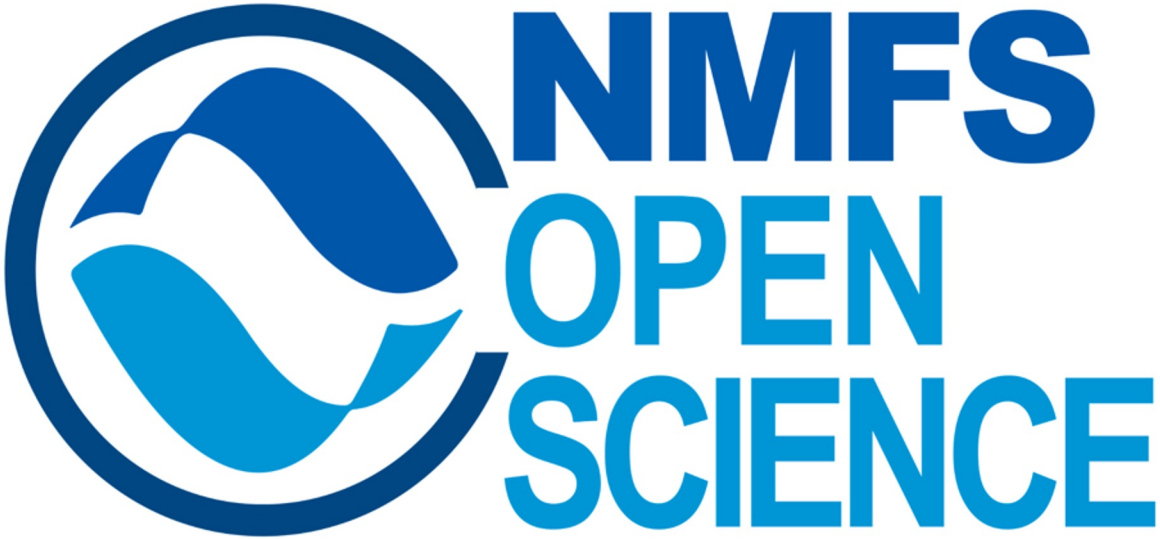
- Python client: pycax on GitHub at nwfsc-math-bio/pycax
- R client: rCAX on GitHub at nwfsc-math-bio/rCAX

## Installation

Development version

```
pip install git+git://github.com/nwfsc-math-bio/pycax.git#egg=pycax-client
```

https://nwfsc-math-bio.github.io/rCAX/


NMFS
OPEN
SCIENCE