Bigelow | Laboratory for Ocean Sciences

Reproducible Research in Ocean Data Science

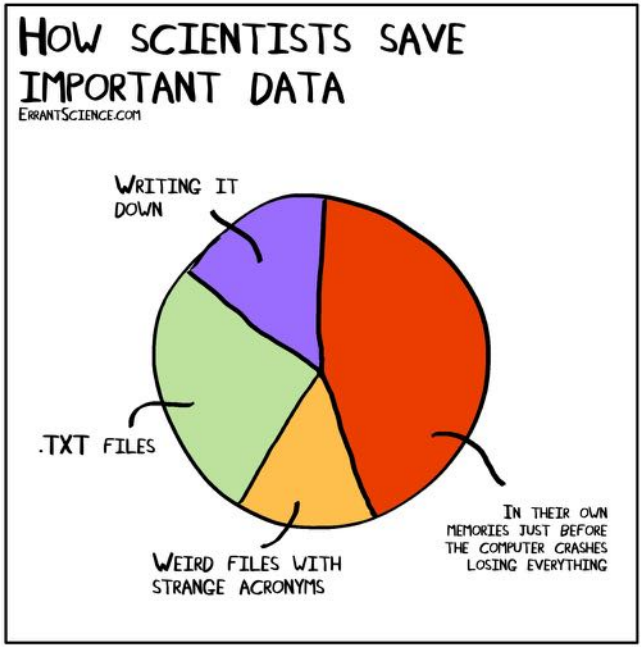
…or: A Gift to Your Future Self

Nick Record

Tandy Center for Ocean Forecasting

OceanHackWeek 2022

So you're working with data, code, and collaborating…



it's a good time to think about *reproducibility*

Resources / Acknowledgements

- **OceanHackWeek 2020: Joseph Gum – *Reproducible Research***
  - https://www.youtube.com/watch?v=qQZVEfljyDI
- **Bigelow Café Code 2021: Cath Mitchell – *Reproducible Research***
  - https://drive.google.com/file/d/1I_Fuz2bNG0rHi8ZH72szMeftNTCJO7I6/view

The Reproducibility Crisis

The Reproducibility Crisis

PLOS MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

The Reproducibility Crisis

- Bias toward publishing statistically significant results
- Negative results lost to filing cabinets(?)
- Documented instances of “cherrypicking” results to get into top journals
- …there have been some high-profile examples (in later slides…)

Reproducibility and Replicability in Science, 2019

- Recently released report from the National Academies of Sciences, Engineering, and Medicine, commissioned by Congress/NSF
- "prompted by concerns about the reproducibility and replicability of scientific research"
- Recommendations on what scientists, funders, journals/conferences, institutions should do in the future
- Recommendations:
  - 6-3: Funding open-source, useable tools and infrastructure for reproducibility
  - 6-5: NSF recommendations related to repositories
  - 6-7: Journals and societies set progressive 3R policies for submissions

National Academies of Sciences, Engineering, and Medicine 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

What is Reproducibility?

Three “R”s of Science

- **Repeatability**: **Same** team, **same** experimental setup
- **Replicability: Different** team, **same** experimental setup
- **Reproducibility: Different** team, **different** experimental setup

- **Methods reproducibility**: provide sufficient detail about procedures and data so that the same procedures could be exactly repeated.
- **Results reproducibility**: obtain the same results from an independent study with procedures as closely matched to the original study as possible.
- **Inferential reproducibility**: draw the same conclusions from either an independent replication of a study or a reanalysis of the original study.

Plesser HE (2018) Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Front. Neuroinform*. 11:76. doi: 10.3389/fninf.2017.00076

Why is it important?

Why is it important?

- Scientific integrity
  - Increased rigor and quality of scientific outputs
  - Greater trust in science

Why is it important?

Retraction Watch

Tracking retractions as a window into the scientific process

https://retractionwatch.com

- Scientific integrity
  - Increased rigor and quality of scientific outputs
  - Greater trust i

- Authors admit to stealing parts of a paper from a thesis on an unrelated subject
- Should residents and fellows be encouraged to publish systematic reviews and meta-analyses?
- How an ivermectin study that didn't mention COVID-19 fell under scrutiny
- 'My egregious delay': Science journal takes more than three years to retract paper after university investigation
- Courage and correction: how editors handle – and mishandle – errors in their journals
- Two abstracts about unapproved heart technology retracted

Why is it important?

- Scientific integrity
  - Increased rigor a
    outputs
  - Greater trust in s

Retraction Watch

Tracking retractions as a window into the scientific process

https://retractionwatch.com

Bad spreadsheet merge kills depression paper, quick fix resurrects it

- The authors of a paper showing a link between immune response and depression requested a retraction after they realized they'd merged two spreadsheets with mismatching ID codes.
- Original conclusion: Lower levels of CSF IL-6 were associated with current depression and with future depression […].
- Revised conclusion: Higher levels of CSF IL-6 and IL-8 were associated with current depression […].

Source: http://retractionwatch.com/2014/07/01/bad-spreadsheet-merge-kills-depression-paper-quick-fix-resurrects-

Why is it important?

Seizure study retracted after authors realize data got "terribly mixed"

From the authors of **Low Dose Lidocaine for Refractory Seizures in Preterm Neonates**:

"The article has been retracted at the request of the authors. After carefully re-examining the data presented in the article, they identified that data of two different hospitals got terribly mixed. The published results cannot be reproduced in accordance with scientific and clinical correctness."

Source: http://retractionwatch.com/2013/02/01/seizure-study-retracted-after-authors-realize-data-got-terribly-mixed/

Retraction Watch

Tracking retractions as a window into the scientific process

https://retractionwatch.com

t merge kills depres-
k fix resurrects it

howing a link between immune
n requested a retraction after they
wo spreadsheets with mismatching

er levels of CSF IL-6 were
epression and with future

ner levels of CSF IL-6 and IL-8
rent depression […].

ch.com/2014/07/01/bad-
spreadsheet-merge-kills-depression-paper-quick-fix-resurrects-

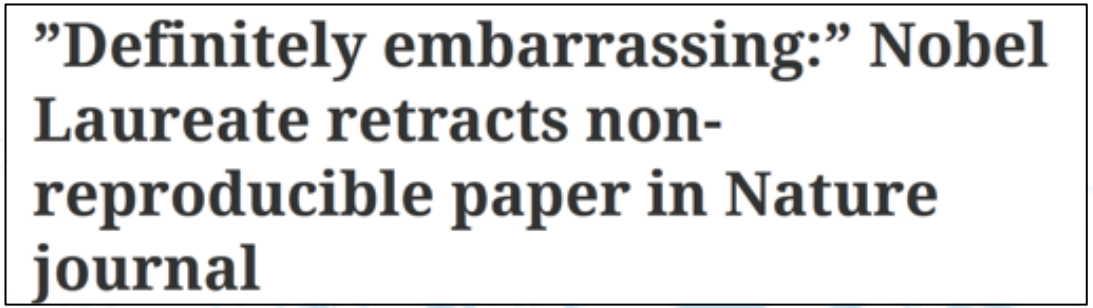
Why is it important?

Retraction Watch

Tracking retractions as a window into the scientific process

https://retractionwatch.com

- Scientific integrity
  - Increased rigor and quality of scientific outputs
  - Greater trust in science



https://retractionwatch.com/2017/12/05/definitely-embarrassing-nobel-laureate-retracts-non-reproducible-paper-nature-journal/

Why is it important?

Retraction Watch

Tracking retractions as a window into the scientific process

https://retractionwatch.com

- Scientific integrity
  - Increased rigor and quality of scientific outputs
  - Greater trust in science

Covid: how Excel may have caused loss of 16,000 test results in England

https://youtu.be/zUp8pkoeMss

"Definitely embarrassing:" Nobel Laureate retracts non-reproducible paper in Nature journal

https://retractionwatch.com/2017/12/05/definitely-embarrassing-nobel-laureate-retracts-non-reproducible-paper-nature-journal/

Why is it important?

- Scientific integrity
  - Increased rigor and quality of scientific outputs
  - Greater trust in science
- Facilitates collaboration

Why is it important?

- Scientific integrity
  - Increased rigor and quality of scientific outputs
  - Greater trust in science
- Facilitates collaboration
- Open science
  - Requirements from funding agencies to share data and code

Reproducibility Activity

Exercise 1 - Part 1

Complete the following tasks and write instructions/documentation for your collaborator to reproduce your work starting with the original dataset

`gapminder-5060.csv`

https://github.com/SeascapeScience/reproducibility/find/main

1. Visualize life expectancy over time for Canada in the 1950s and 1960s using a line plot.
   Stretch goal: Add lines for Mexico and US.

2. Visualize the relationship between GDP and life expectancy for countries in Europe in 1952.
   Stretch goal: Add a line for 1967 in another color.

Exercise 1 - Part 2

Introduce yourself to your collaborator and tell them why you're here.

1. Swap instructions/documentation with your collaborator, and try to reproduce their work without talking to each oher. If your collaborator does not have the software they need to reproduce your work, we encourage you to either help them install it or walk them through it on your computer in a way that would emulate the experience.

2. Then, talk to each other about challenges you faced (or didn't face) or why you were or weren't able to reproduce their work.

Exercise 1 - Wrap up

- Have you ever tried to reproduce someone else's data analysis before?
- Have you ever tried to reproduce your own work before?
- What tools did you use and were you successful in reproducing your collaborator's work?
- What made it easy/hard for reproducing your parners' work?
- What would have to happen if you had to extend the analysis further?
- If you caught a data error how easy/hard would it be to re-create the analysis?
- What would happen if your collaborator is no longer available to walk you through their analysis?

Reproducibility Best(?) Practices

Best Practices

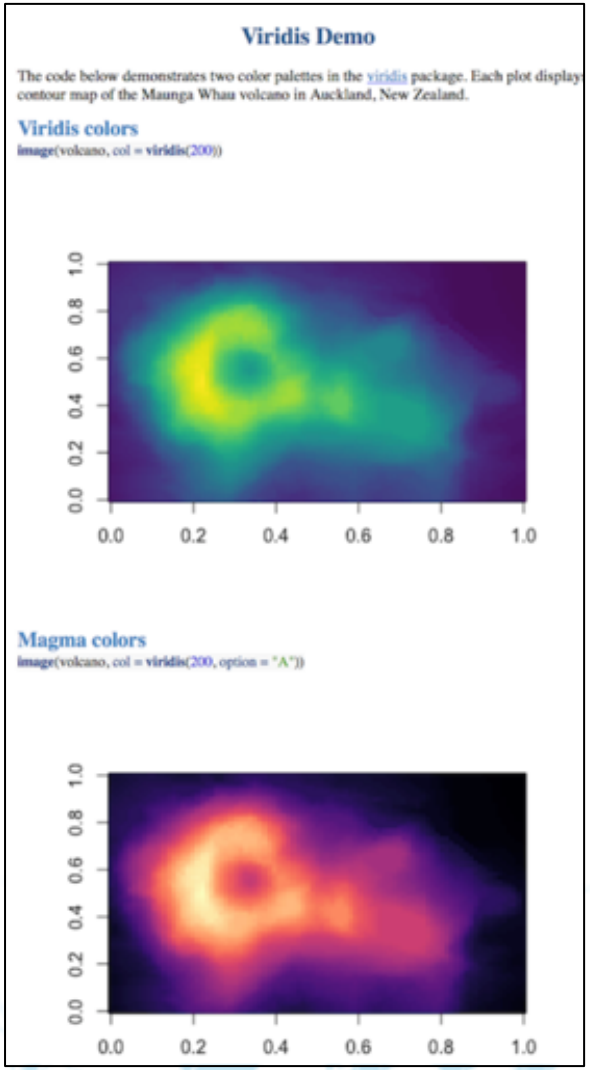
1. Documentation
2. Organization
3. Automation
4. Dissemination

Documentation

Document everything!

Developed protocols/workflows, what you did today, any figures you created...

1. Electronic lab books
    a. All-singing-all-dancing: expensive, but comprehensive
    b. Simple note-taking: e.g. Evernote, OneNote (Windows)
    c. Interactive computing e.g. RMarkdown, Jupyter notebooks
2. Text documents
    a. Simple text editors e.g. notepad
    b. Word processors e.g. Word, Google Docs



Organization

Think about file organization **before** you start your project

Three points to keep in mind:

1. Consistent, computer readable, human understandable names (for files, spreadsheet names, etc)
    a. No spaces (use _ or -)
    b. Do you want to sort files? Consider numbering them, or using dates in the format YYYYMMDD or YYYY.MM.DD
2. Make it easy to find data, code, documents, figures
    a. Both by **future you** and a **colleague**
3. Find something that works for you: there's no right answer here

Organization - Example directory structure

Expressive Project	Non Expressive Project
/01-scripts 01-process-landsat-data.py 02-calculate-ndvi.py 03-create-ndvi-maps.py /02-data /raw-data ← /landsat-imagery /june-2016 /july-2016 /cold-springs-fire-boundary /03-output-graphics ndvi-map-june-2016.png ndvi-map-july-2016.png /04-final-paper veg-impacts-cold-springs-fire.pdf	work.py plotting.py plotting-test.py landsat/ data-file.txt old-stuff/ testoutput1.txt testoutput2.csv

Organization - Example directory structure

ProjectName/

- README.MD
- Dataset/
  - Raw Data/
  - Processed Data/
    - YYYY-MM-DDVersion
    - YYYY-MM-DDVersion
- Analysis (or Code)/
  - Data cleaning/
  - Data preprocessing/
  - Output/
    - Graphs
    - Tables
- Publications/
  - .tex files
  - .bib file

Automation

- Think about developing methods as small, reusable chunks
  - e.g. for cleaning data, making figures, doing analysis, running a type of simulation
- Select the best tool for the job you are trying to do
- If you're working with spreadsheets:
  - Think about OpenRefine, it helps for working with messy data, and tracks what you're doing
  - Have a **README** file that goes along with your Excel document
  - **Document** how you are creating your figures
- If you're coding:
  - Try using **functions**, particularly for tasks you do a lot of times
  - Have an **individual script to make a figure** - try creating all the details of the figure within your script so you don't have to post-edit in e.g. Photoshop

Dissemination

- Typical ways we share our research: journal articles, poster presentations, oral presentations
  - Use reference managers for keeping track of citations - really useful!
    - e.g. EndNote, Mendeley, Zotero
  - Make it easy on yourself to redo figures if you need to tweak your analysis
- Other things to think about:
  - Often, we need to share our data on online repositories
  - For our research to be truly reproducible, our methods need to be shared e.g. share our scripts or developed workflows

Takeaways:

1. Document, document, document

2. Organize your project directories to make them easier to understand

3. Make things simple on yourself - *“Your closest collaborator is you six months ago, but you don’t reply to email”*

4. Find a system that works for you, your project, your research group

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, Eivind Hovig

Published: October 24, 2013 • https://doi.org/10.1371/journal.pcbi.1003285

- For Every Result, Keep Track of How It Was Produced
- Avoid Manual Data Manipulation Steps
- Archive the Exact Versions of All External Programs Used
- Version Control All Custom Scripts
- Record All Intermediate Results, When Possible in Standardized Formats
- For Analyses That Include Randomness, Note Underlying Random Seeds
- Always Store Raw Data behind Plots
- Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Connect Textual Statements to Underlying Results
- Provide Public Access to Scripts, Runs, and Results

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285

Online resources

https://www.earthdatascience.org/courses/intro-to-earth-data-science/open-reproducible-science/get-started-open-reproducible-science/best-practices-for-organizing-open-reproducible-science/

https://guides.lib.berkeley.edu/c.php?g=652220&p=4575532

https://kbroman.org/Tools4RR/assets/lectures/06_org_eda_withnotes.pdf

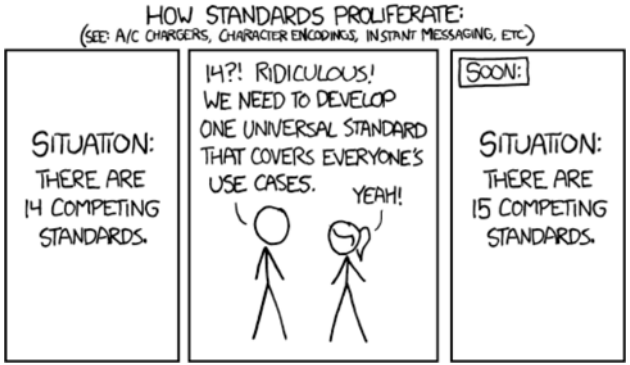
https://openrefine.org/

Some Specifics for Ocean Data Science

Selected slides from: Joseph Gum, with modifications
https://github.com/asx-/reproducible-science

Quiz time!

- What is a repository?
  - Differences between code and data repositories?
- Who do you turn in your data/results to?
  - Is your code publicly accessible?
- Name a version control software (VCS)
- What is a branch in git?
- What is a package in Python, R?
- What are FAIR principles?
- What are CARE principles?
- Have you had trouble running code from others, or sharing your code?
- Is interoperability important to you?
  - Do you feel there are too many standards to understand?



Standards - xkcd.com

FAIR Principles

- Make data easier to use: “guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets”
- **What does FAIR mean?**
  - **Findable**
  - **Accessible**
  - **Interoperable**
  - **Reusable**

- “Is your data FAIR?” vs “Does your data follow FAIR virtues?”
  - Absolutes versus relatives – start with the relatives

https://www.go-fair.org/fair-principles/

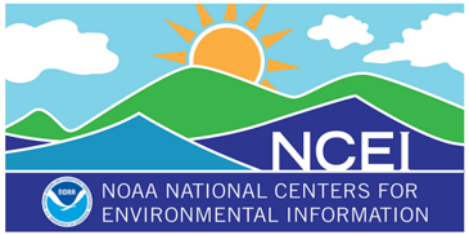
CARE Principles

- *The **CARE Principles for Indigenous Data Governance*** were developed by the Global Indigenous Data Alliance (GIDA) in 2019 to complement the FAIR principles and other movements towards Open Data.
- **What does CARE mean?**
  - **Collective benefit**
  - **Authority to control**
  - **Responsibility**
  - **Ethics**

https://atlanticdatastream.ca/en/article/fair-and-care-data-principles

Data Centers/Repositories

- Data centers are paid to make your data available – reach out to them
  - (It's also a requirement of your grant to submit your data)
- Use data center libraries to read our data holdings, or ask them to make libraries
  - Feel free to contribute too!



(Oceanography) Data Stewards Salad

- ESIP
- RDA
- EarthCube
- FAIR
- Ocean Best Practices
- DataONE
- IEDA
- Force11
- And so on…

Repository of Methods – Ocean Best Practices

- DOIs for manuals and handbooks
  - How to do fieldwork, make models, etc.
- UNESCO/IOC project: “Ocean Best Practices”
- Internationally agreed upon methods…
  - …and publically submitted methods
- Journals: Best Practices in Ocean Observing, Deep Sea Research, etc.

United Nations Educational, Scientific and Cultural Organization
Intergovernmental Oceanographic Commission
Ocean best practices

https://www.oceanbestpractices.org/

Version Control: git, Github

- Easier to keep track of yours/others changes
- Remote copies as failsafes in case of computer failure – but not a replacement for backups

- Commits should be “atomic”
  - Commits are fully formed, not half-done
  - Related edits are grouped into a single commit
  - Commits are small (one change) and often
- Use branches to try new ideas without polluting the master
- Want more git tutorials?

- Github – one of many places hosting code repos

"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com

Resources for improving (digital) reproducible science

- Carpentries Project
  - Software
  - Data
- OceanHackWeek
  - 2018 - …

Year of Open Science

National Aeronautics and Space Administration

NASA

TOPS

A NASA OPEN-SOURCE SCIENCE MISSION:
TOPS: TRANSFORM TO OPEN SCIENCE

Dr. Chelle Gentemann, TOPS Program Scientist
Yvonne Ivey, TOPS Project Manager
Cyndi Hall, TOPS Community Coordinator
Isabella Martinez, TOPS Curriculum Coordinator
Dr. Yaitza Luna-Cruz, OSSI/TOPS Science Coordinator

Kevin Murphy, Chief Science Data Officer SMD
Katie Baynes, Deputy Chief Science Data Officer SMD
Dr. Steve Crawford, Science Data Officer SMD
Dr. Elena Steponaitis, Program Officer
Amy (Uyen) Truong, Chief Science Data Office Coordinator
Christian Reyes, OSSI Coordinator

Shelley Stall, Vice President, Data Leadership, AGU
Lauren Parr, Senior Vice President, Meetings & Learning, AGU
Chris Erdmann, Assistant Director, Data Stewardship,AGU
Laura Lyon, Program Manager, Science, AGU
Brooks Hanson, Executive Vice President, Science, AGU

Thanks!