

# Data and Doughnuts

March, 14th 2025



# Agenda - morning 09:00 a.m. - 12:00 a.m.

Scientific Introduction (Harrison)

Data handling and publishing

- Data Handling (Johanna)
- Nice examples from Justin and Jarod
- FAIR data and data publishing (Harrison)
- Dataverse (Johanna)
- Nice example from Jahred

Study design and statistical analysis

Data analysis

- How to work with spreadsheets (Harrison)
- What are jupyter notebooks and hands-on exercises (Johanna)



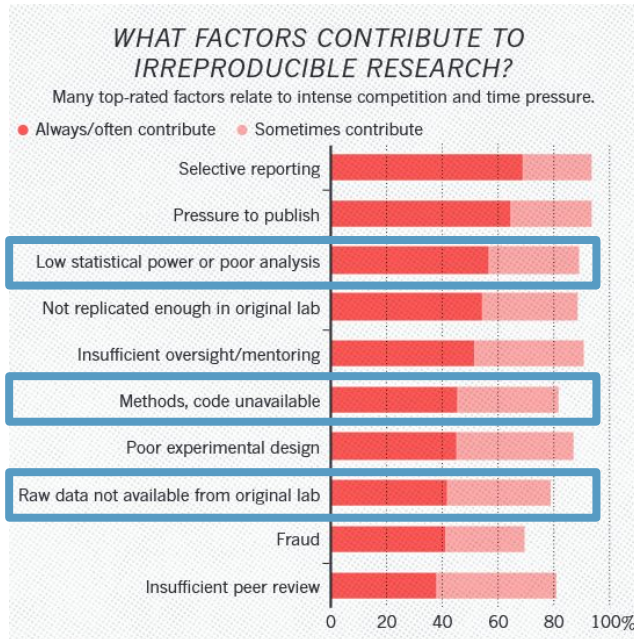
# Agenda - afternoon 01:00p.m.-02:30p.m.

Lunch (12:00-01:00 p.m.)

Open session (01:00 p.m. - 02:30 p.m.)

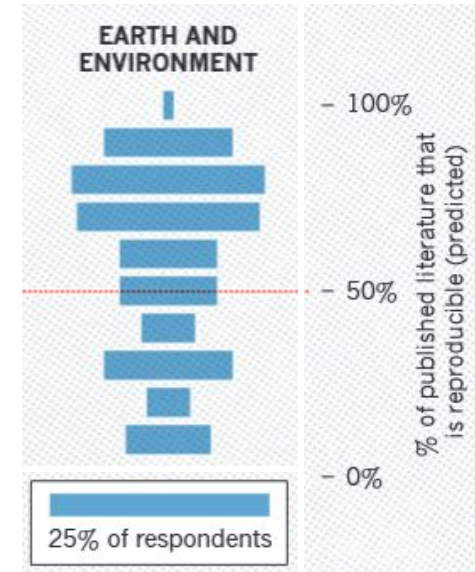
- Open questions
- Hands on exercises

# THE UNDERLYING PROBLEM



More than 25 % of 1576 researchers asked in 2016 thought that more than 50% of published science in environmental research was not reproducible.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility.

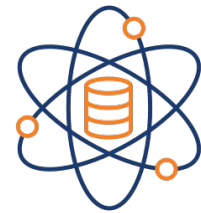


# BEYOND REPRODUCIBILITY: PROMOTING REUSE



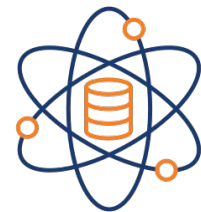
In 2016, the [FAIR Guiding Principles for scientific data management and stewardship](#) (Wilkinson et al) were published in *Scientific Data*. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Source: <https://www.go-fair.org/fair-principles/>



# Data Handling

Johanna

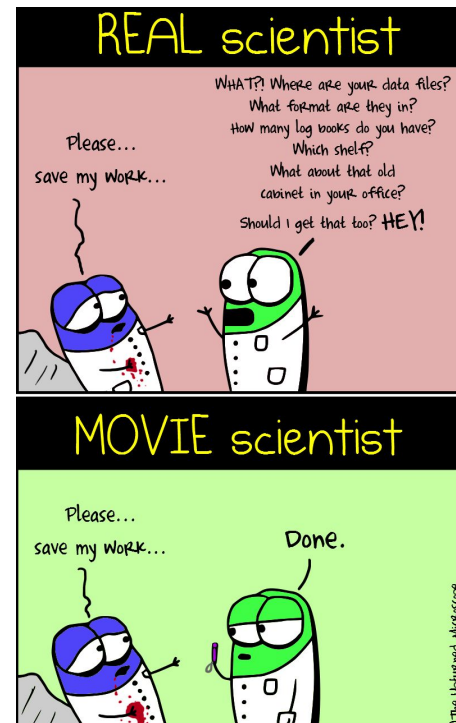


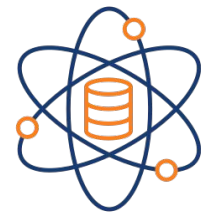
# Goal: have all necessary data available when writing up your PhD thesis

And how?

- **Delete** 'useless files', **archive** relevant data.
- Use meaningful **folder names and structures**
- Use **filenames** that help you navigate your data.
- **Document** your work and data.

“Reproducibility is like brushing your teeth. It is good for you, but it takes time and effort. Once you learn it, it becomes a habit.” - Irakli Loladze, mathematical biologist





# Archive relevant data

- Get rid of data and files, not everything is important.
- Archive your important data.
- **Raw experimental data** / simulation inputs are important data!
- Keep track, take notes, and document.
  - **Digitize** and keep your **lab notes**.
- **Backup** your data.







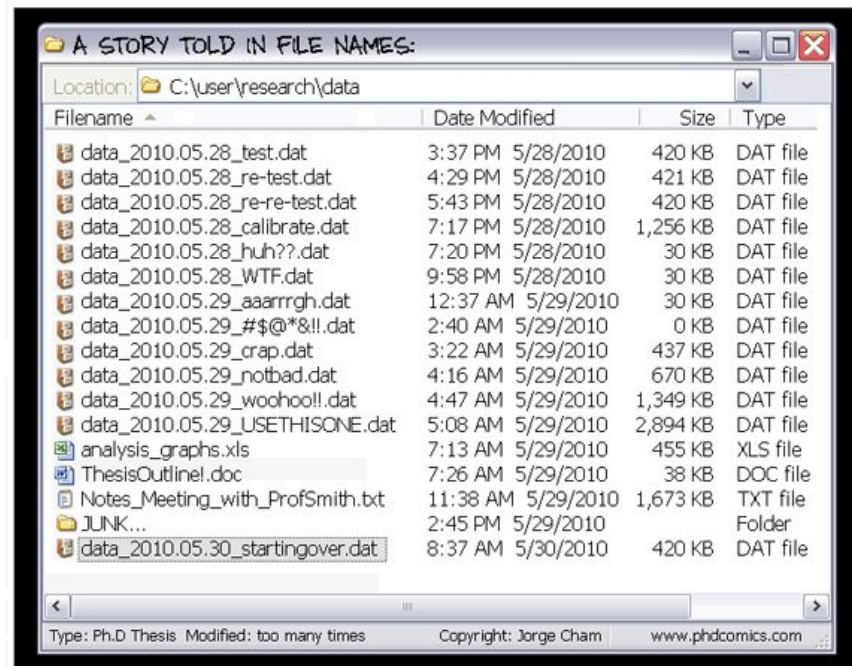
# Folder names and structures

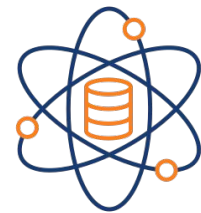
- What important contextual information can you use to sort data files?:
  - separate data by academic semester?
  - separate data by projects?
  - separate data by measurement or instrument type?
  - ....
- Make a system. Follow the system. Be consistent.
  - **dedicated places for raw data and lab notes**
  - **dedicated places for data analysis/code**
  - ...
- Where is your data saved? (special care with myDrive and OneDrive)



# Filenames

- Avoid white spaces.
- Exclude special characters:  
& , \* % # ; \* ) ( ! @ \$ ^ ~ ' { [ ? < - .
- Period only before file extension.
- 32 characters or less.
- numeric versioning: \_v001, \_v002
- ISO 8601 format for dates:  
20240101, 2024\_01\_01





# File naming conventions

- What **information is important** about your files and **makes each file distinct**?
- Choose common naming convention and **stick to it!**

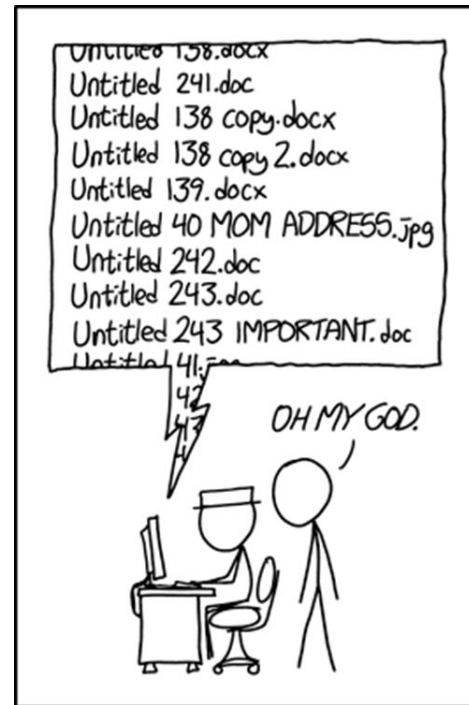
Nice Example

DOC_WAX_01	SeaW_WAX_01	pH4_WAX_01	pH8_WAX_01_S
DOC_WAX_02	SeaW_WAX_02	pH4_WAX_02	pH8_WAX_02_S
DOC_DEX_01	SeaW_DEX_01	pH4_DEX_01	pH8_DEX_01_S
DOC_DEX_02	SeaW_DEX_02	pH4_DEX_02	pH8_DEX_02_S
DOC_COL_01	SeaW_COL_01	pH4_COL_01	pH8_COL_01_S
DOC_COL_02	SeaW_COL_02	pH4_COL_02	pH8_COL_02_S

Sample

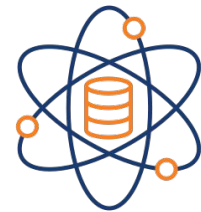
Sorbent

Number



<http://imgs.xkcd.com/comics/documents.png>

# KEEP YOUR DATA, AND KEEP IT CLEAN!

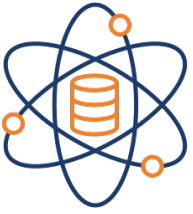


- **Delete** 'useless files', **archive** relevant data.
- Use meaningful **folder names and structures**
- Use meaningful **file names**.
- **Document** your work and data.



<https://pub-e93d5c9fdf134c89830082377f6df465.r2.dev/2024/06/Messy-Chaotic-Data.webp>

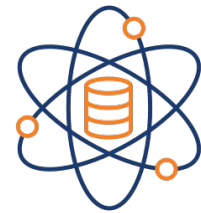
# RESSOURCES



- Worksheet: Naming and organizing your files and folders (MIT libraries):  
[https://www.dropbox.com/scl/fi/1zd63iszw33rh4h1cu1dl/Worksheet\\_fileOrg.docx?rlkey=q0t25t1wtp4qx2p1ne39qfhd&dl=0](https://www.dropbox.com/scl/fi/1zd63iszw33rh4h1cu1dl/Worksheet_fileOrg.docx?rlkey=q0t25t1wtp4qx2p1ne39qfhd&dl=0)
- File naming conventions worksheet (Caltech Libraries):  
<https://authors.library.caltech.edu/records/mmpnf-cez11>
- Google Drive for Desktop  
[Google Drive for Desktop](#)
- One Drive  
[OneDrive](#)



<https://pub-e93d5c9fdf134c89830082377f6df465.r2.dev/2024/06/Messy-Chaotic-Data.webp>



# Data Handling

Justin, Jarod



# Data Publishing

Harrison



# Five Key Aspects of Data Publishing to Keep in Mind

1. Data Management and Documentation
2. Choosing the Right Repository
3. Licensing and Access Control
4. Citations and Persistent Identifiers
5. Compliance and Ethical Considerations

Understanding these aspects ensures that data publishing adds value to the academic community, boosts research impact, and upholds ethical standards. Let's explore them a little further.





# Data Management and Documentation

Proper data organization, metadata creation, and clear documentation (e.g., methods, formats, and variables) are essential to ensure that others can understand and reuse the dataset. Following standards like the FAIR principles enhances data clarity and usability.



# Choosing the Right Repository

Selecting an appropriate repository — whether general-purpose (e.g., Dataverse, Zenodo, Figshare) or discipline-specific (e.g., GenBank for genetic data) — ensures long-term preservation, proper indexing, and greater visibility for the dataset.



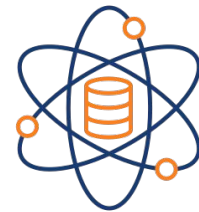
# Licensing and Access Control

Applying the right license (e.g., Creative Commons) clarifies terms of use and reuse. Researchers should also determine access levels, balancing openness with ethical or legal obligations related to sensitive data.



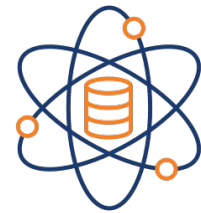
# Citations and Persistent Identifiers

Assigning persistent identifiers like DOIs (Digital Object Identifiers) makes the dataset citable, ensuring proper attribution and enabling other researchers to reference the data in future work.



# Compliance and Ethical Considerations

Researchers must comply with institutional, funder, and journal data-sharing policies while addressing ethical concerns, such as obtaining consent for data sharing and protecting sensitive or personal information.



# Harvard Dataverse

Johanna



# Harvard Dataverse Login

- Go to: <https://dataverse.harvard.edu>
- Login with your SSO

## ➡ Log In


Log in or sign up with your institutional account — more information about account creation. Leaving your institution? Please contact Harvard Dataverse Support for assistance.

Your Institution

University of Rhode Island  
University of Rhode Island

University of Rhode Island ▼ Continue

University of Padova  
University Of Passau  
University of Pennsylvania  
University of Peradeniya  
University of Perugia  
University of Pilsen - Test  
University of Pisa  
University of Pittsburgh  
University of Politecnica delle Marche  
University of Porto  
University of Pretoria  
University of Pretoria TEST  
University of Redlands  
University of Regina  
University of Rhode Island  
University Of Rochester  
University of Ruhuna (Wellamadama Site)  
University of Salerno

**HARVARD**  
Dataverse

[Add Data](#) ▼ [Search](#) ▼ [About](#) [User Guide](#) [Support](#) [Sign Up](#) [Log In](#)

Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

[Add a dataset +](#)

Organize datasets and gather metrics in your own repository.

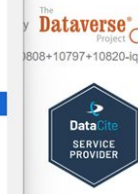
A dataverse is a container for all your datasets, files, and metadata.

[Add a dataverse +](#)

Publishing your data is easy on Harvard Dataverse!

Learn about getting started creating your own dataverse repository here.

[Getting started](#)



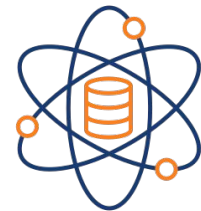
# Harvard Dataverse

... is a free repository for research data (and code)

...is commonly used in research  
( > 75k datasets published)

...'**general repository**' - might not  
be the best choice for specific  
data sets (e.g. raw .mzML)

... assigns a DOI to each  
published data set



Metrics

87,943,120 Downloads

Advanced Search

+ Add Data

Dataverses (1)

Datasets (8)

Files (7)

**Dataverse Category**

Research Project (1)

**Metadata Source**

Harvard Dataverse (12)  
Harvested (4)

**Publication Year**

2024 (6)  
2023 (2)  
2022 (3)  
2021 (1)

**License**

CC0 1.0 (10)  
CC BY-NC-SA 4.0 (1)

**Subject**

Earth and Environmental Sciences (6)  
Medicine, Health and Life Sciences (3)  
Engineering (2)  
Chemistry (1)  
Social Sciences (1)

**Author Name**

Liddie, Jahred (3)  
Andvik, Clare (1)  
Becanova, J. (1)  
Blind, Marie-Abèle (1)  
Elsie Sunderland (1)

More...

1 to 10 of 16 Results

Sources, Transport, Exposure and Effects of PFAS (University of Rhode Island)  
Oct 1, 2024  
 Collection of research data generated in the superfund research program STEEP. This includes (i) quantification of PFAS concentrations in environmental samples (water, air, sediment, biota, etc.) to study bioaccumulation of PFAS in the food web and analyze passive sampling methods for PFAS detection in means of targeted mass spectrometry, (ii) metabolism studies to study compound effects of PFAS exposure and diet on the health of mice and bacteria, and (iii) metadata of cohort studies on PFAS effects on the Faroe islands.

**PFAS Statewide Sampling Dataset**  
Dec 14, 2023  
 Liddie, Jahred, 2022, "PFAS Statewide Sampling Dataset", <https://doi.org/10.7910/DVN/8LPLCE>, Harvard Dataverse, V3  
... This repository contains a representative sample of U.S. community water systems (CWS) included in the PFAS statewide sampling dataset. In the first versions of this dataset, 18 states were included. The statewide sampling dataset now compiled data from 24 U.S. statewide sampling campaigns of CWS for per- and polyfluoroalkyl substances (PFAS) in drinking water. ...  
Subtitle: PFAS drinking water sampling data from U.S. community water systems  
Keyword Term: Drinking water, PFAS, community water systems

**Massachusetts PFAS Influence Map Project**  
Sep 3, 2024  
 Hui, Alice, 2024, "Massachusetts PFAS Influence Map Project", <https://doi.org/10.7910/DVN/B8JD7A>, Harvard Dataverse, V2, UNF:6:LFbXdq3AdZoUW/JI+oDcQ== [fileUNF]  
... PFAS exposure has been linked to adverse health effects, including developmental issues in children, reduced fertility, and increased risk of certain cancers. Understanding where PFAS contamination exists is crucial for protecting both human health and the environment. ...  
Related Publication URL: <https://alicexhui.github.io/pfas/>  
Keyword Term: PFAS

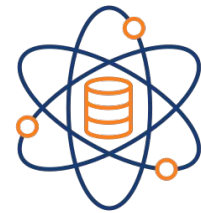
**Characterizing the Areal Extent of PFAS Contamination in Fish Species Downgradient of AFFF Source Zones**  
Sep 6, 2024  
 Pickard, Heidi, 2024, "Characterizing the Areal Extent of PFAS Contamination in Fish Species Downgradient of AFFF Source Zones", <https://doi.org/10.7910/DVN/PICNIV>, Harvard Dataverse, V1  
... Dataset of targeted PFAS and extractable organofluorine concentrations in water, sediment, and biological tissue samples that support data included in two papers: "Characterizing the Areal Extent of PFAS Contamination in Fish Species Downgradient of AFFF Source Zones" and "Bioaccumulation of Perfluoroalkyl Sulfonamides (FASA)".  
Notes: This dataset contains concentration data for targeted PFAS analytes and extractable organofluorine (EOF) in surface water, sediment, and biological tissue samples from sites within an AFFF-impacted watershed. The dataset additionally includes information on coordinates for each site,





# Data Publishing in Harvard Dataverse

Jahred Liddie



# Best practices with spreadsheets

Harrison



# Keeping track of your analyses

When you're working with spreadsheets, during data clean up or analyses, it's very easy to end up with a spreadsheet that looks very different from the one you started with. In order to be able to reproduce your analyses or figure out what you did when Reviewer #3 asks for a different analysis, you should:

- create a new file with your cleaned or analyzed data. Don't modify the original dataset, or you will never know where you started!
- keep track of the steps you took in your clean up or analysis. You should track these steps as you would any step in an experiment. We recommend that you do this in a plain text file stored in the same folder as the data file.

# Spreadsheet setup example



survey\_data.xlsx

Home Layout Tables Charts SmartArt Formulas Data

N10

	A	B	C	D	E	F	G	H	I
1	DateCollected	Year	Month	Day	Plot	Species	Sex	Weight	
2	7/16/13	2013	7	16	2	DM	F		
3	7/16/13	2013	7	16	7	DM	M	33g	
4	7/16/13	2013	7	16	3	DM	M		
5	7/16/13	2013	7	16	1	DM	M		
6	7/18/13	2013	7	18	3	DM	M	40g	
7	7/18/13	2013	7	18	7	DM	M	48g	
8	7/18/13	2013	7	18	4	DM	F	29g	
9	7/18/13	2013	7	18	4	DM	F	46g	
10	7/18/13	2013	7	18	7	DM	M	36g	
11	7/18/13	2013	7	18	7	DM	F	35g	

2013-clean 2014-raw 2014-clean

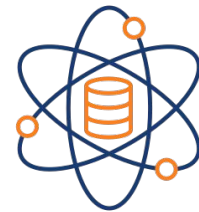
README\_surveyData.txt — Carpe... Add License

Processing notes on survey\_data.xlsx

2014-08-19 work done -----

1. Transferred 2013-raw to 2013-clean, and 2014-raw to 2014-clean
2. In 2013-clean: created a 'Species' column and moved information from header to that column
3. In 2013-clean, put all the different tables together into one table with columns: date collected, plot, species, sex, weight
4. In 2013-clean, separated month/day/year column into three columns for year, month, and day using formulas

Line: 11:94 Plain Text Tab Size: 4



# Structuring data in spreadsheets

The cardinal rule of using spreadsheet programs for data is to keep it “tidy”:

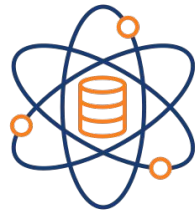
1. Put all your variables in columns - the thing you’re measuring, like ‘weight’ or ‘temperature’.
2. Put each observation in its own row.
3. Don’t combine multiple pieces of information in one cell. Sometimes it just seems like one thing, but think if that’s the only way you’ll want to be able to use or sort that data.
4. Leave the raw data raw - don’t change it!
5. Export the cleaned data to a text-based format like CSV (comma-separated values) format. This ensures that anyone can use the data, and is required by most data repositories.



# Exercise - Data Cleaning

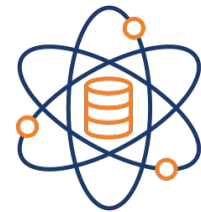
1. Download the data by clicking [here](#) to get it from FigShare.
2. Open up the data in a spreadsheet program (e.g. Excel or Google Sheets).
3. You can see that there are two tabs. Two field assistants conducted the surveys, one in 2013 and one in 2014, and they both kept track of the data in their own way in tabs 2013 and 2014 of the dataset, respectively. Now you're the person in charge of this project and you want to be able to start analyzing the data.
4. With the person next to you, identify what is wrong with this spreadsheet. Also discuss the steps you would need to take to clean up the 2013 and 2014 tabs, and to put them all together in one spreadsheet.

Important: Do not forget our first piece of advice: to create a new file (or tab) for the cleaned data, never modify your original (raw) data.



# Introduction to jupyter notebooks

Johanna



# Introduction to jupyter notebooks

- **interactive computing environment:**

- Load and cleanup data
- Execute data analysis
- Visualize data
- **Document all steps**

- **Why jupyter notebook?**

- reproducible
- automated workflow
- helpful packages and features

## Show off example

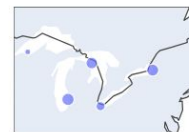
Mapping in jupyter notebook is easy. In the code block below you find three code lines used to show PFAS interactive map. You might have to zoom in.

```
[70]: import plotly.express as px

fig = px.scatter_geo(fish_data_pfos_mean, lat='Latitude', lon='Longitude',
                    hover_name="Lake", size=selected_compound,
                    animation_frame="Year",
                    scope='north america',
                    title="Great Lakes")

fig.show()
```

Great Lakes

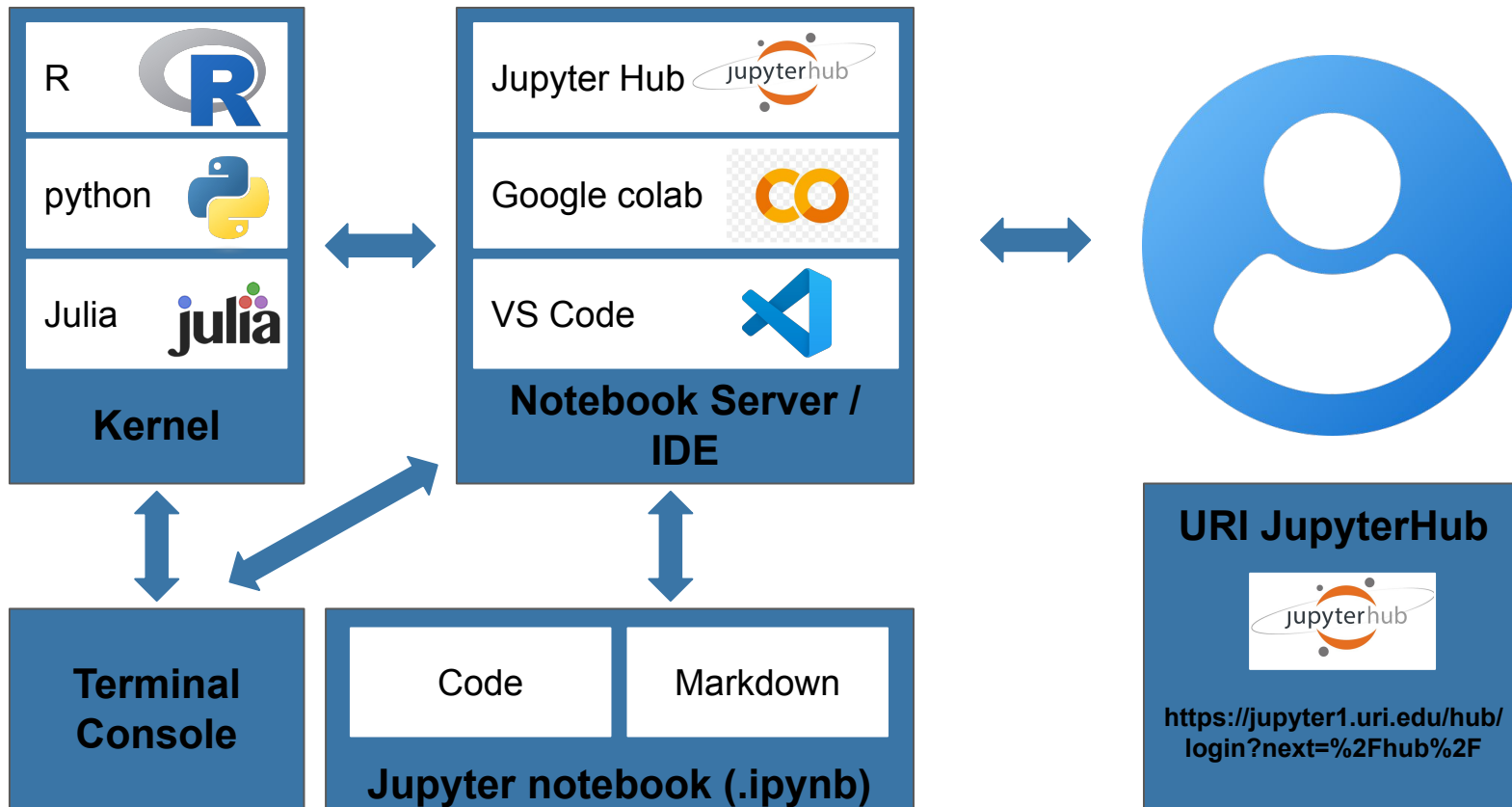


Year=2022

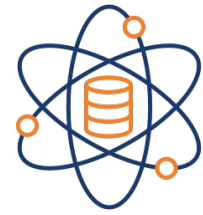




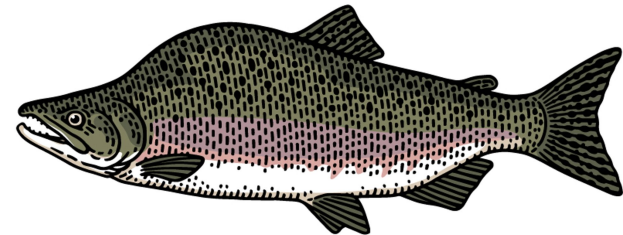
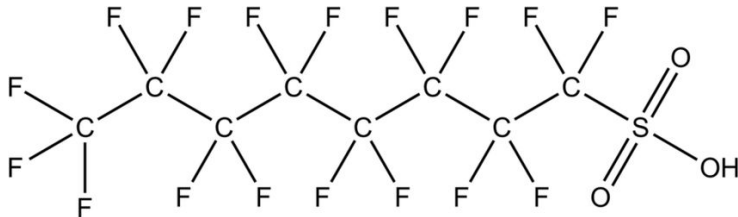
# Jupyterhub - relevant terms

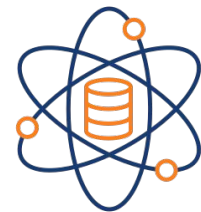


# Example of today



- Great Lakes Environmental Database
- PFAS concentrations of fish caught at 10 stations at the great lakes from 2011 - 2022.
- Goal: Read in data, filter data, average data, plot data,...
- **Notebook and example data provided in your jupyterhub account**





## Further resources

- (1) basic introduction: <https://www.dataquest.io/blog/jupyter-notebook-tutorial/>
- (2) pandas: [https://pandas.pydata.org/docs/getting\\_started/intro\\_tutorials/index.html](https://pandas.pydata.org/docs/getting_started/intro_tutorials/index.html)
- (3) matplotlib: [https://www.w3schools.com/python/matplotlib\\_plotting.asp](https://www.w3schools.com/python/matplotlib_plotting.asp)
- (4) General data science: [https://github.com/engineersCode/EngComp1\\_offtheground](https://github.com/engineersCode/EngComp1_offtheground)

and many more...

Your account for URI's jupyter lab will be open **until the end of summer semester**. Download all relevant work beforehand and contact [johanna.ganglbauer@uri.edu](mailto:johanna.ganglbauer@uri.edu), if you want to use it in the long term.