RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Lecture 1: Introduction

16:375:501 Environmental Science Analysis
Fall 2024, Department of Environmental Sciences

Instructor: Xiaomeng Jin (xiaomeng.jin@rutgers.edu)

Class Format

- This is an online asynchronous course, meaning that we do not “meet”, not even via the web.
- You decide when to do the work.
- To prevent you from procrastinating too much, you will have an assignment due each week for the first 12 weeks.
- Each week, you will need to:
  - Watch a short video from the instructor to introduce the key concepts and things to do for this week.
  - Self-study the materials provided (Jupyter Notebook, videos, readings, website).
  - Complete a quiz that is either embedded in the videos or reading materials.
  - Follow the instructions to finish the weekly assignment.

An example of your weekly assignment

Pros and Cons of Online Course

- **Flexibility**

You can study at your own pace and on your own schedule, making it easier to balance with other commitments

- **Accessibility**

Allows you to access courses from anywhere.

- **Limited interaction**

Less face-to-face interaction with instructors and peers.

- **Self-Discipline required**

Requires strong time-management skills and self-motivation to stay on track

An example of data analysis pipeline in environmental sciences

1. Conceptualize problem and question

What are the impacts of wildfires on air quality?

2. Formulate hypothesis and predictions

Wildfires cause increasing air pollution episodes.

3. Design experiment or develop methods

Statistical approach or modeling approach.

4. Collect data

Wildfire data, air quality measurements, model data.

5. Process and analyze data

Python, statistical analysis.

6. Present results and discussion of findings in broader context

Figures, publications, presentations

Challenges in Environmental Science Analysis

- **Complexity**

The things we need to do are very complicated and hard!

- **Data Quality & Quantity**

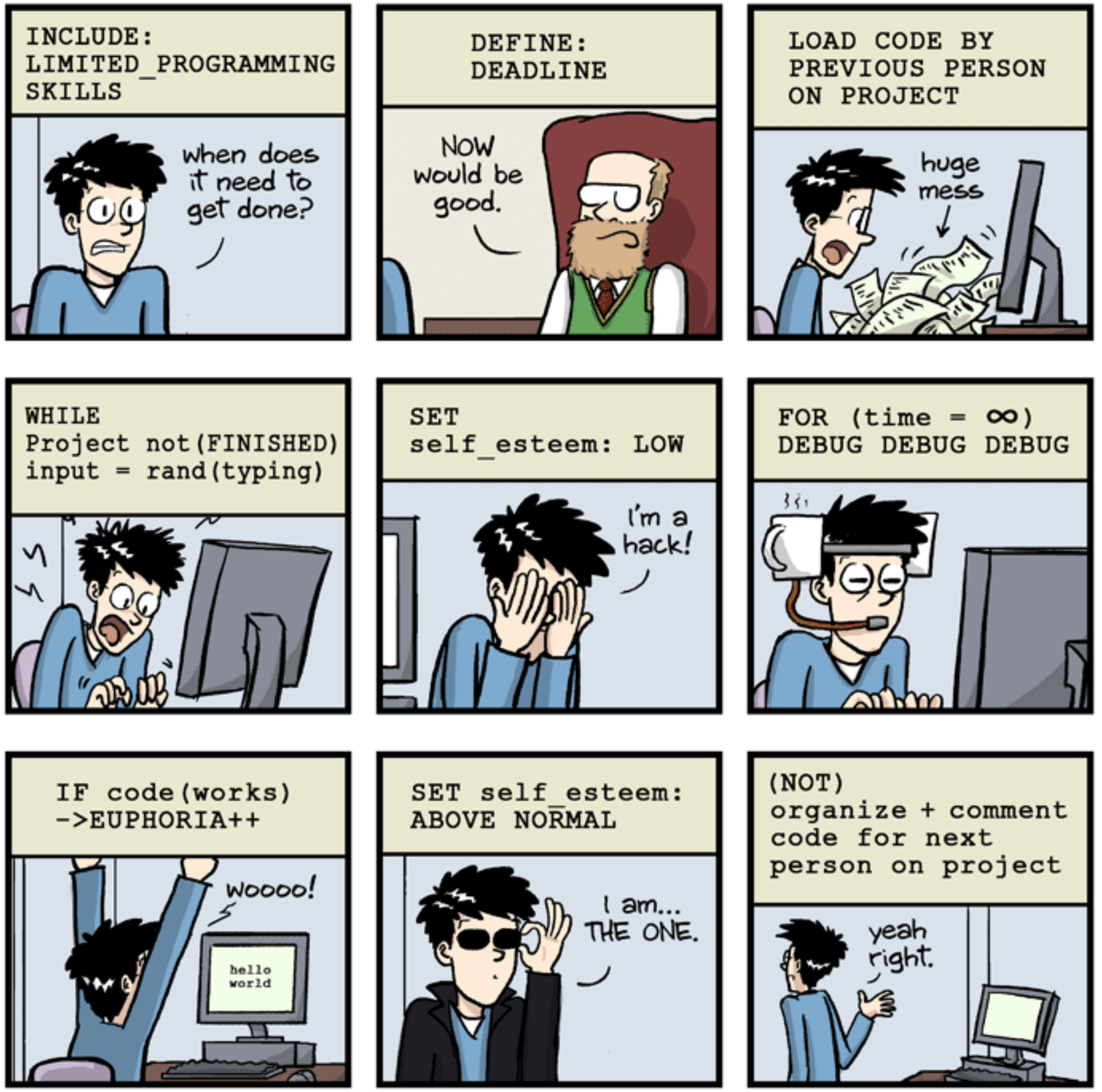
Environmental datasets are often incomplete with errors.

- **Reproducibility**

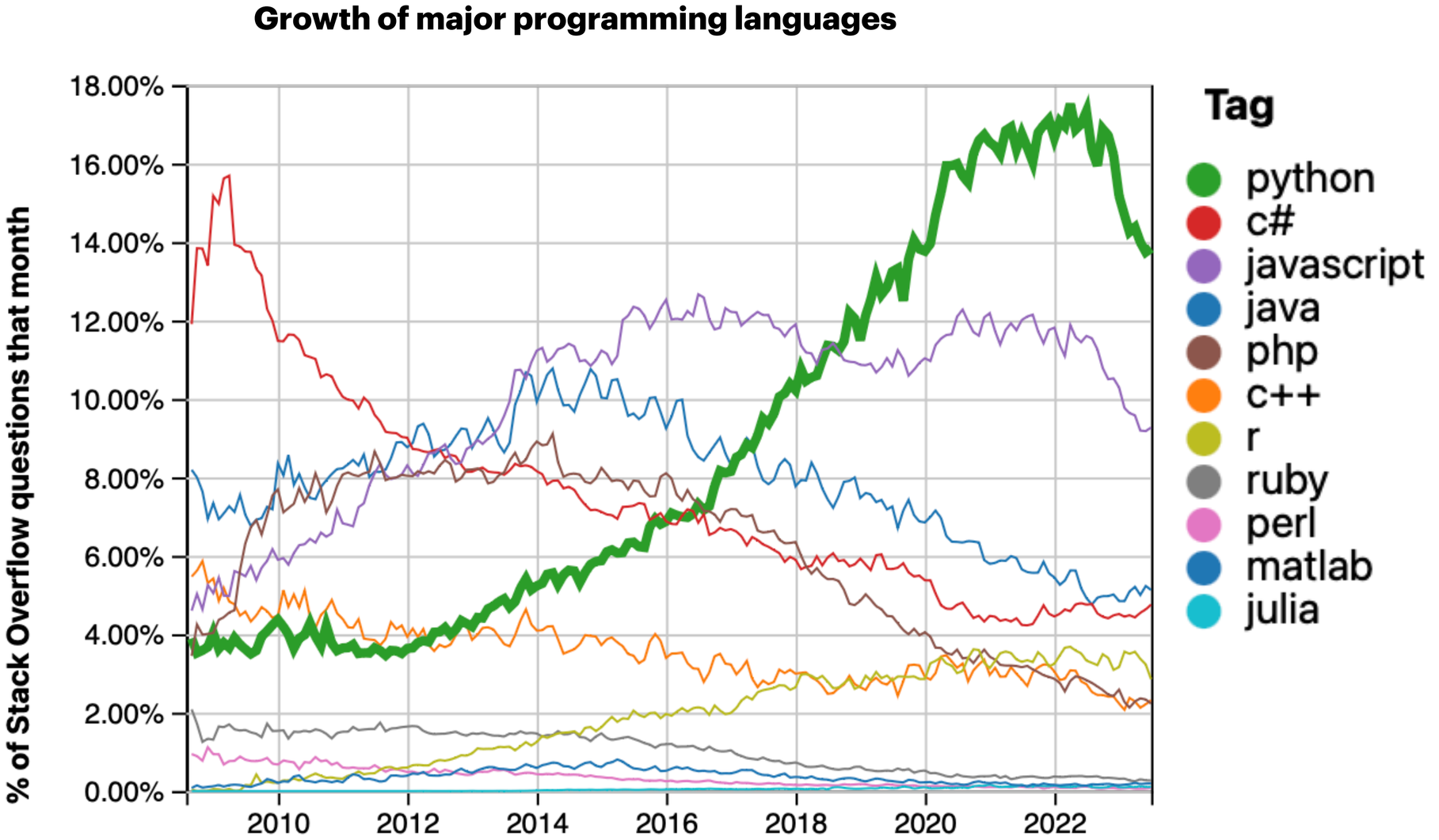
It’s not easy to reproduce others’ work (or even our own).

Problem: Complexity
Solution: Write as little code as possible!

PROGRAMMING FOR NON-PROGRAMMERS

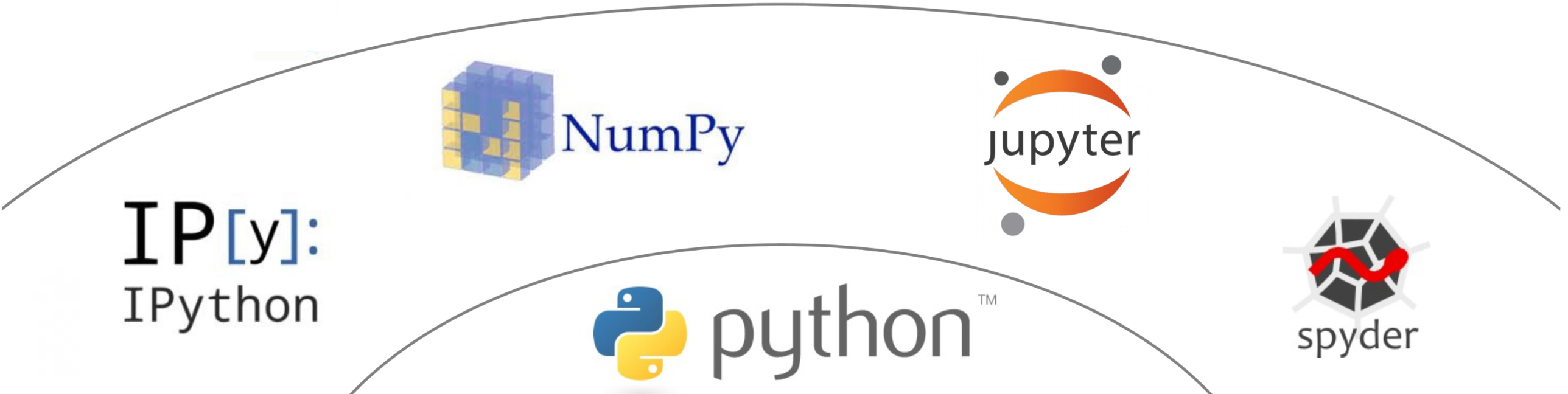


A high-level programming language with a healthy, active community and a broad range of packages

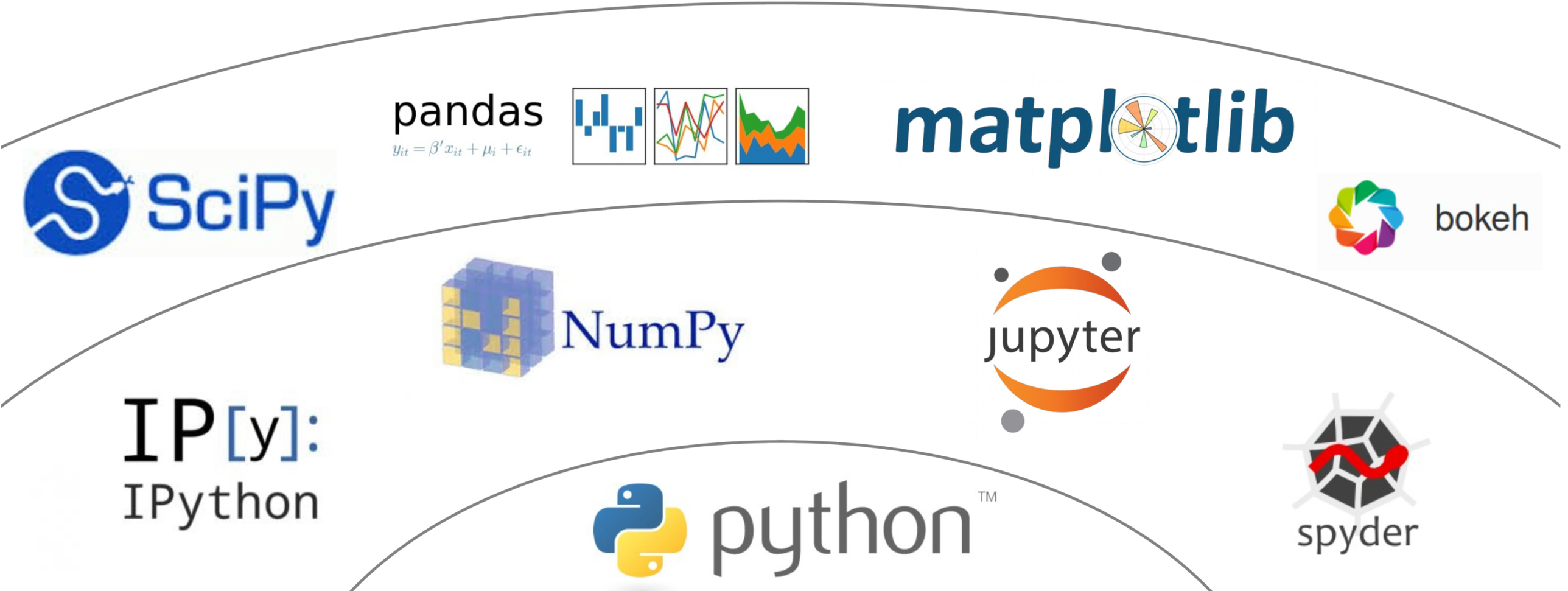


Scientific Python Ecosystem

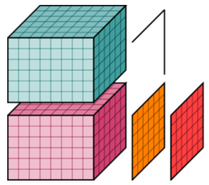
Scientific Python Ecosystem



Scientific Python Ecosystem



Scientific Python Ecosystem



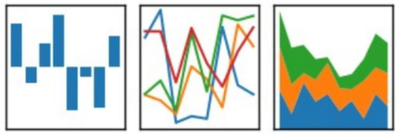
xarray

GeoPandas

MetPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



matplotlib

bokeh

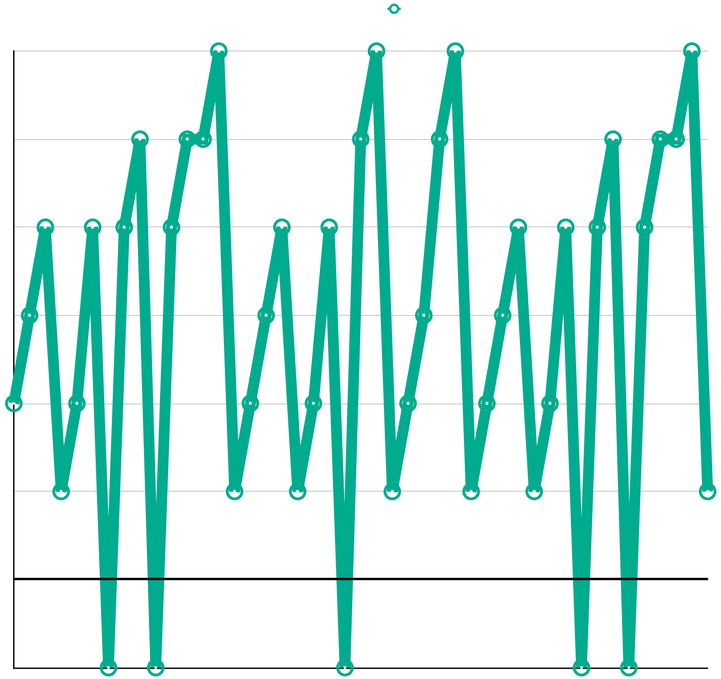
IP[y]:
IPython

python™

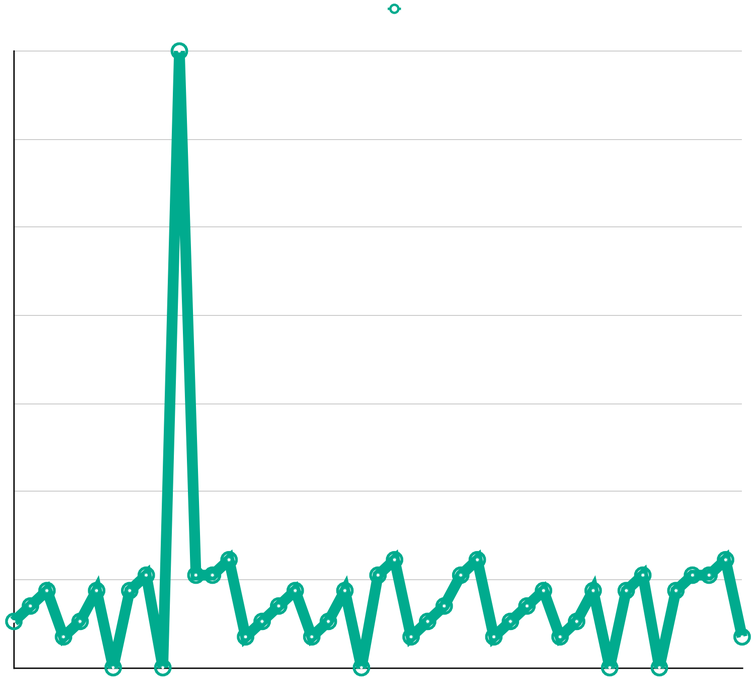
Problem: Data Quality and Quantity
Solution: Learn effective statistical methods to interpret environmental datasets.

Common data problems

Non-detects



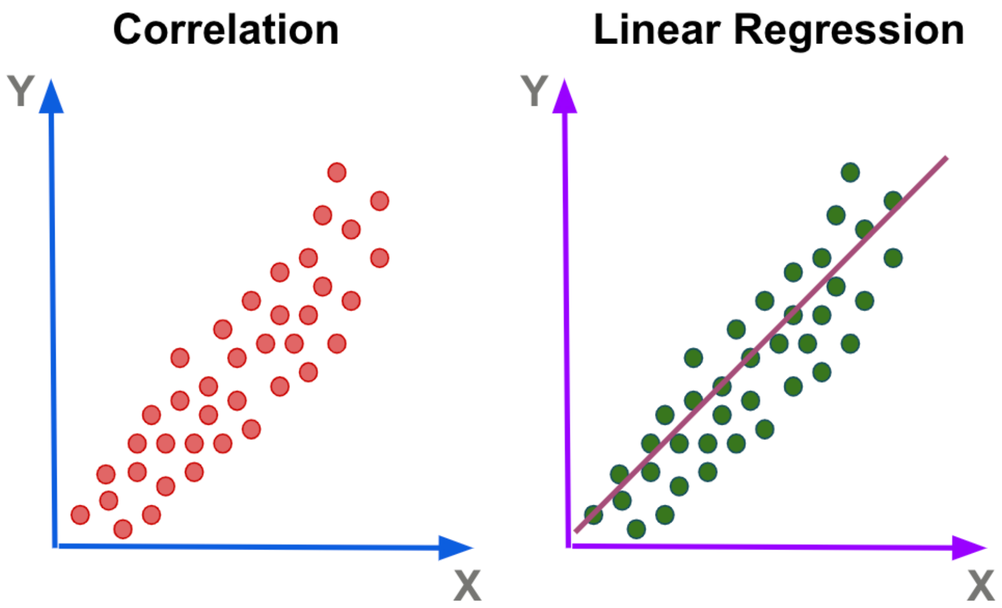
Data Anomaly



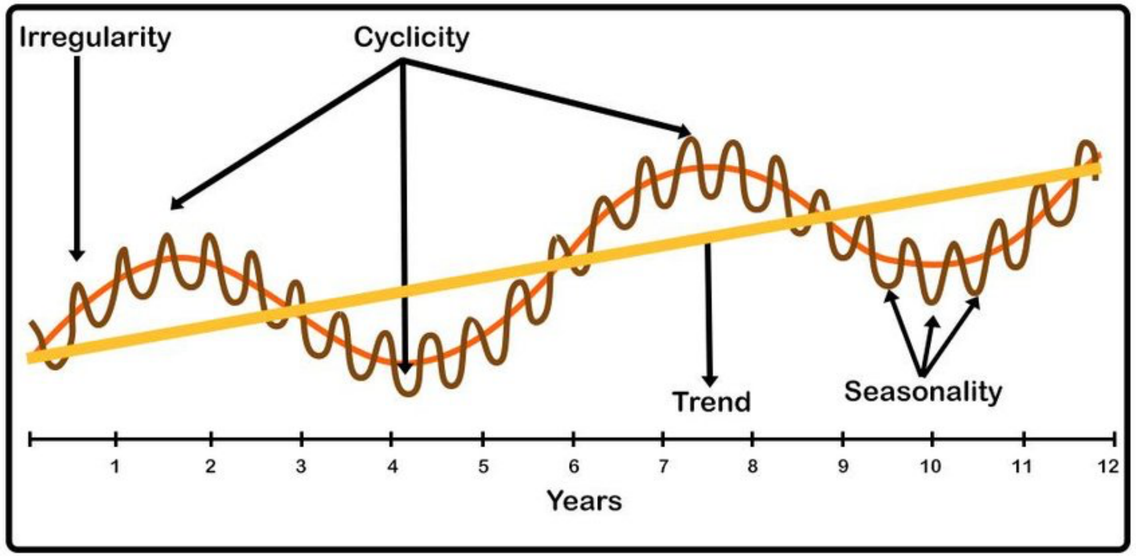
Missing Data

Common Statistical Methods for Environmental Science Analysis

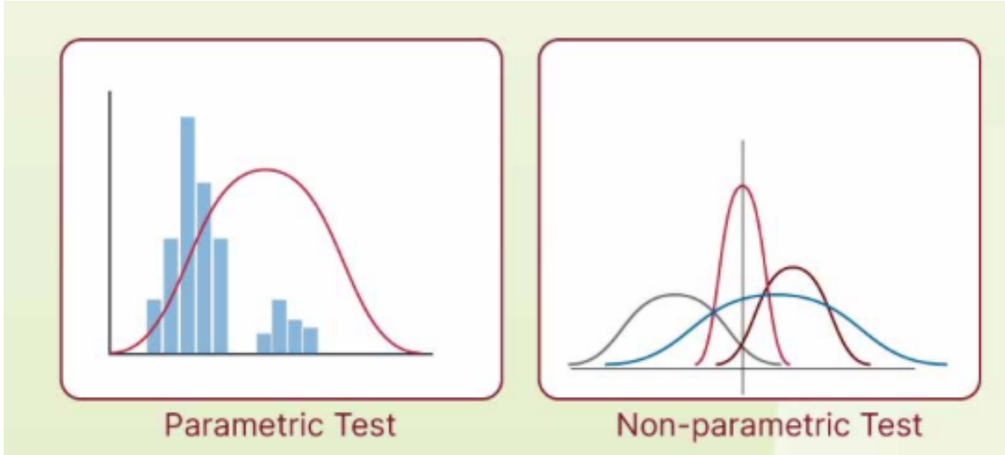
Correlation and Regression



Time series analysis



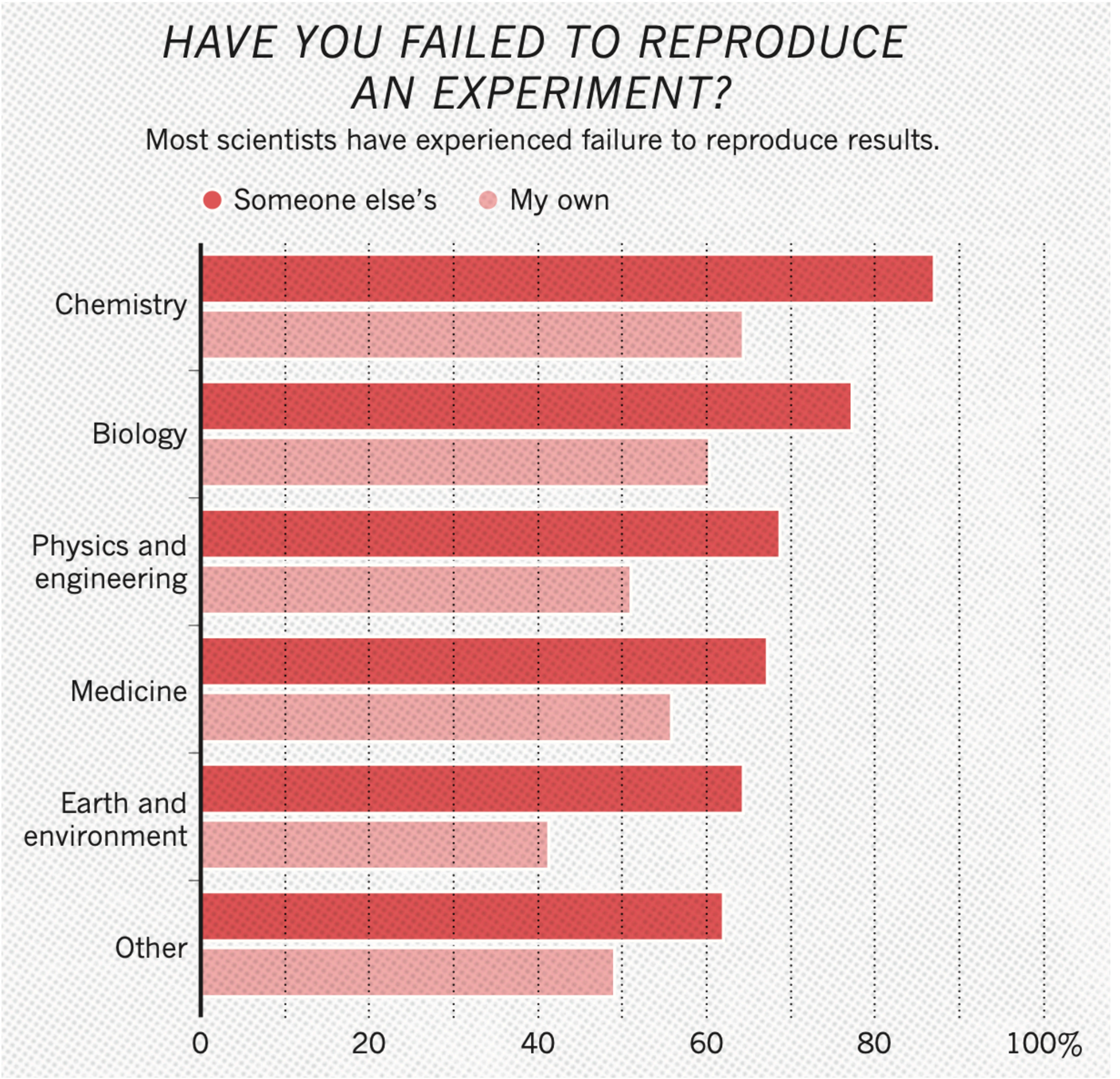
Hypothesis test



Problem: Reproducibility
Solution: Learn and embrace open-science practices

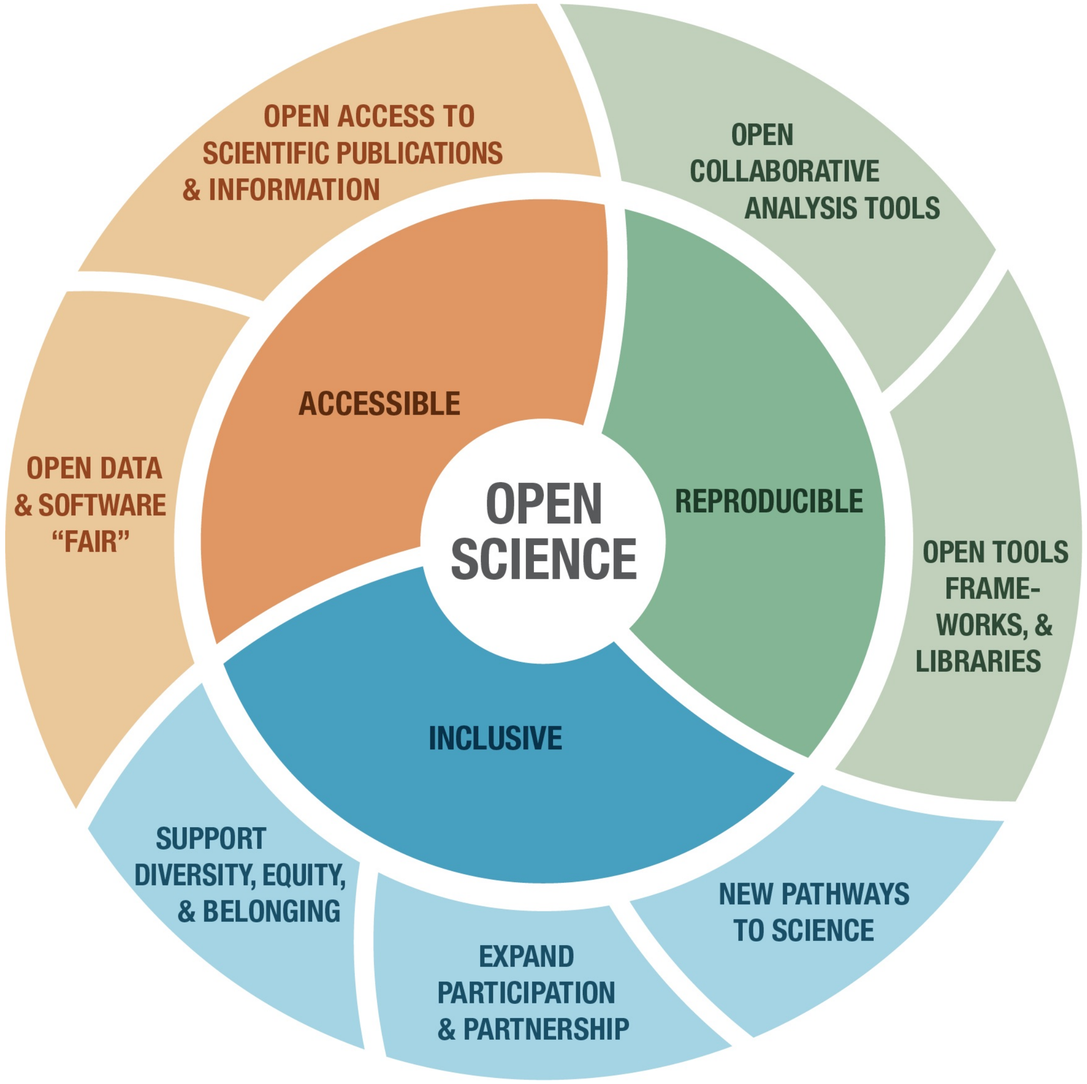
What is reproducibility?

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable



Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016). https://doi.org/10.1038/533452a

The Open Science Vision



Reproducible: The scientific process and results should be open such that they are reproducible by members of the community.

Accessible: Data, tools, software, documentation, publications should be findable, accessible, interoperable and reusable (FAIR) to all.

Inclusive: The process and participants should welcome participation by and collaboration with diverse people and organizations.

https://www.earthdata.nasa.gov/technology/open-science/oss-for-eso-workshops

NASA, Open-Source Science Initiative, Kevin Murphy, 2021

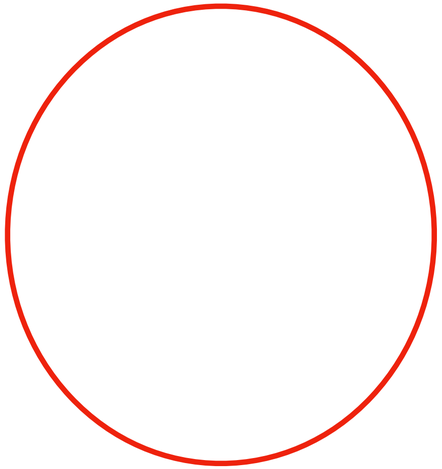
What topics will we cover in class?

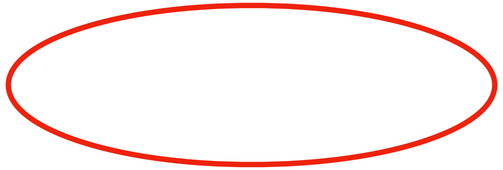
- Introduction to programming in the open-source Python language and in-depth exploration of the numerical analysis and visualization packages that comprise the modern scientific Python ecosystem.
- Introduction to statistical methods for environmental data processing, analysis and interpretation.
- Introduction to the concept of open science and best practices for conducting reproducible environmental science analysis.

What you should accomplish by the end of semester?

1. A collection of Jupyter notebooks with tutorials and examples that you can refer to.
2. A collection of analytical methods that you know how to use in your future work or research.
3. A final project that showcases your ability to conduct environmental science analysis.

Course Website:





Schedule

Grading

- Weekly Quiz: 20%
- Weekly Assignment: 50%
  - Assignments will be posted on Sunday every week.
  - Due the next Monday at 12 PM.
- Final Project: 30%

Assignment 1: Due Sept 9

1. Introduction about yourself:
   - Your name, major, prior data analysis experience.
   - What are the challenges you have in doing environmental science analysis?
   - What you want to learn from this course?
2. Preparation for the course:
   - Mark this course on your calendar: block out at least 3 hours each week for this class.
   - Create a folder named ‘environmental_science_analysis’. Save all your future lecture materials in the folder.
3. Final project topic:
   - Think about a environmental science topic you’re interested.