Rensselaer

why not change the world?®

Data (and Data Science) in Scientific Research

Ahmed Eleish

Data Science ITWS/CSCI/ERTH-4350/6350

October 28th, 2026

Tetherless World Constellation
Rensselaer Polytechnic Institute



Contents

- Data lifecycle and research reproducibility
- Open science and research data sharing
- Research infrastructure and collaboration

01 | Data Lifecycle and Research Reproducibility



https://uwyo.libguides.com/publishing-research-data

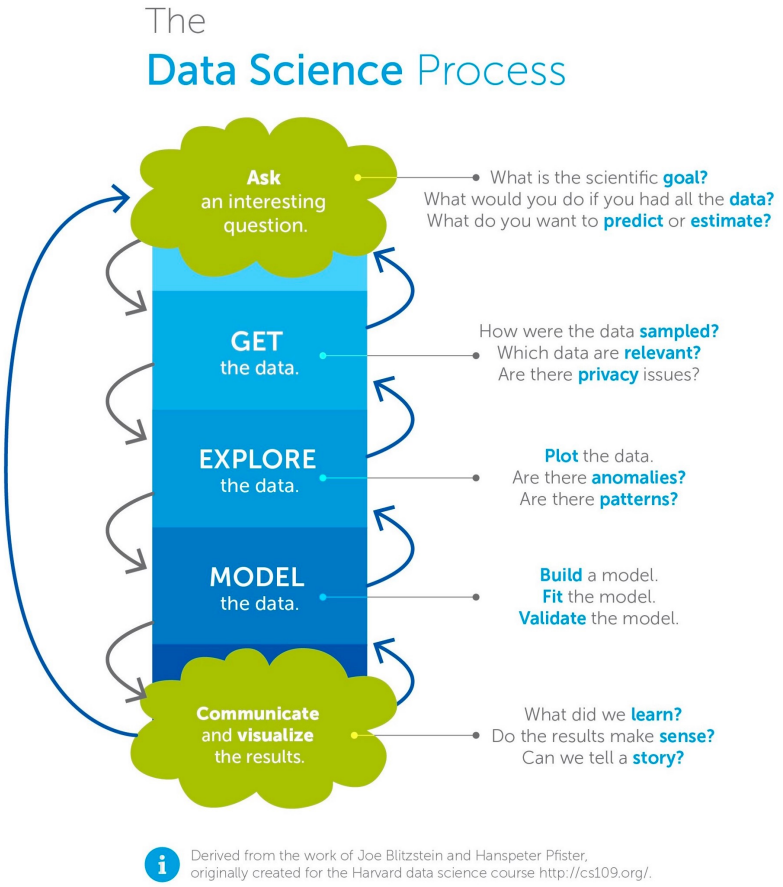


Image Credit: Joe Bilizsten, Hanspeter Pfister / Harvard Data Science

The Scientific Method…

- Define a question
- Gather information and resources (observe)
- Form an explanatory hypothesis
- Test the hypothesis by performing an experiment and collecting data in a reproducible manner
- Analyze the data
- Interpret the data and draw conclusions that serve as a starting point for a new hypothesis
- Publish results
- Retest (frequently done by other scientists)

https://en.wikipedia.org/wiki/Scientific_method

Primary, Secondary, Observational, Experimental

- Primary data are collected for the current study; secondary data were collected for another purpose.
- Observational studies measure without assigning the treatment.
- Experiments assign conditions to support causal comparisons.
- Each design supports different claims and assumptions.

Primary, Secondary, Observational, Experimental

Dimension	Type	Example	What it supports / key caveat
Data source	**Primary**	Researchers collect new survey responses for a specific question	Can be tailored to the question; collection takes time and resources
Data source	**Secondary**	Researchers analyze an existing census, archive, or published dataset	Efficient and may offer large samples; original measures or collection context may not fit
Study design	**Observational**	Researchers record exposures and outcomes without assigning a treatment	Can reveal patterns and associations; confounding may limit causal conclusions
Study design	**Experimental**	Researchers assign participants or units to treatment and comparison conditions	Can support causal conclusions, especially with random assignment; generalizability still matters

What Makes an Analysis Reproducible?

- Code, data, dependencies, and instructions must be available as permitted.
- Software versions, configuration, and random seeds should be recorded where relevant.
- A computational workflow should be rerunnable from inputs to results.
- Reproducibility includes the research decisions, not just the code script.

Cleaning Without Losing the Evidence

- An untouched raw-data copy must be kept along with any processed version separately described.
- Exclusions, imputations, unit conversions, scale transformations, etc. should be recorded.
- Cleaning decisions can change estimates and scientific conclusions.



Provenance and Metadata

- Provenance records where data came from and how they changed along the way..
- Metadata describe variables, units, methods, dates, etc. in the collection context.
- A data dictionary/ controlled vocabulary makes definitions explicit across a research team/ research community.
- Good documentation lets others assess fitness for reuse.



Versioning the Analysis

- Version control should be used to track code and documentation changes.
- Datasets and analysis releases should be given stable versions or identifiers.
- Raw data, intermediate products, and final outputs should be separate.
- Commit messages and change logs should explain why work changed.

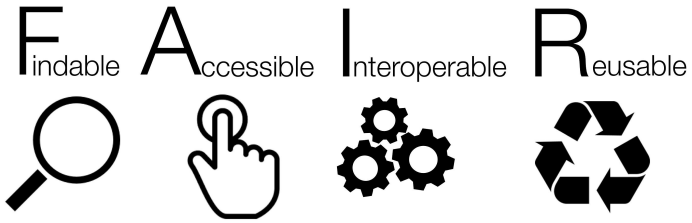
Computational Reproducibility Checklist

- Including a README, data dictionary, and environment specification.
- Documenting steps to obtain inputs and rerun the analysis.
- Making outputs traceable to code and input versions.
- Considering reproducibility tenets for computational workflows.

02 | Open Science and Research Data Sharing

FAIR Data

- Findable: persistent identifiers and searchable metadata.
- Accessible: clear retrieval protocols, including controlled access where needed.
- Interoperable: shared formats and vocabularies enable combination.
- Reusable: sufficient context, licensing, and provenance support new use.



FAIR Does Not Always Mean Public

- FAIR describes data management, not unrestricted publication.
- Sensitive data may require authentication or a trusted access process.
- Metadata can be openly discoverable while data remain controlled.
- Access decisions should align with consent, law, and research risk.

Findability Through Metadata

- Use descriptive titles, keywords, methods, and coverage information.
- Persistent identifiers such as DOIs help cite and locate datasets over time.
- Metadata should match community standards where available.
- A repository record makes data discoverable beyond the original project.





Iris

Donated on 6/30/1988

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.

Dataset Characteristics
Tabular

Subject Area
Biology

Associated Tasks
Classification

Feature Type
Real

 # Instances
150

 # Features
4

Dataset Information

What do the instances in this dataset represent?
Each instance is a plant

Additional Information
This is one of the earliest datasets used in the literature on classification methods and widely used in statistics and machine learning. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other....

SHOW MORE

Has Missing Values?
No

DOWNLOAD (3.7 KB)

IMPORT IN PYTHON

CITE

352 citations

1558221 views

Keywords

ecology

Creators

R. A. Fisher

DOI

10.24432/C56C76

License

This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



Choosing an Access Model

- Public access supports broad reuse when risks are low.
- Restricted access can protect confidentiality or contractual obligations.
- Controlled access uses review, agreements, and secure environments.
- Choose the **least restrictive** model consistent with responsibilities.

Consent, Licensing, and Data-Use Agreements

- Consent should anticipate sharing and future use where feasible.
- Licenses specify permitted reuse and attribution conditions.
- Data-use agreements define safeguards and limits for restricted data.
- Check whether third-party rights or terms limit redistribution.

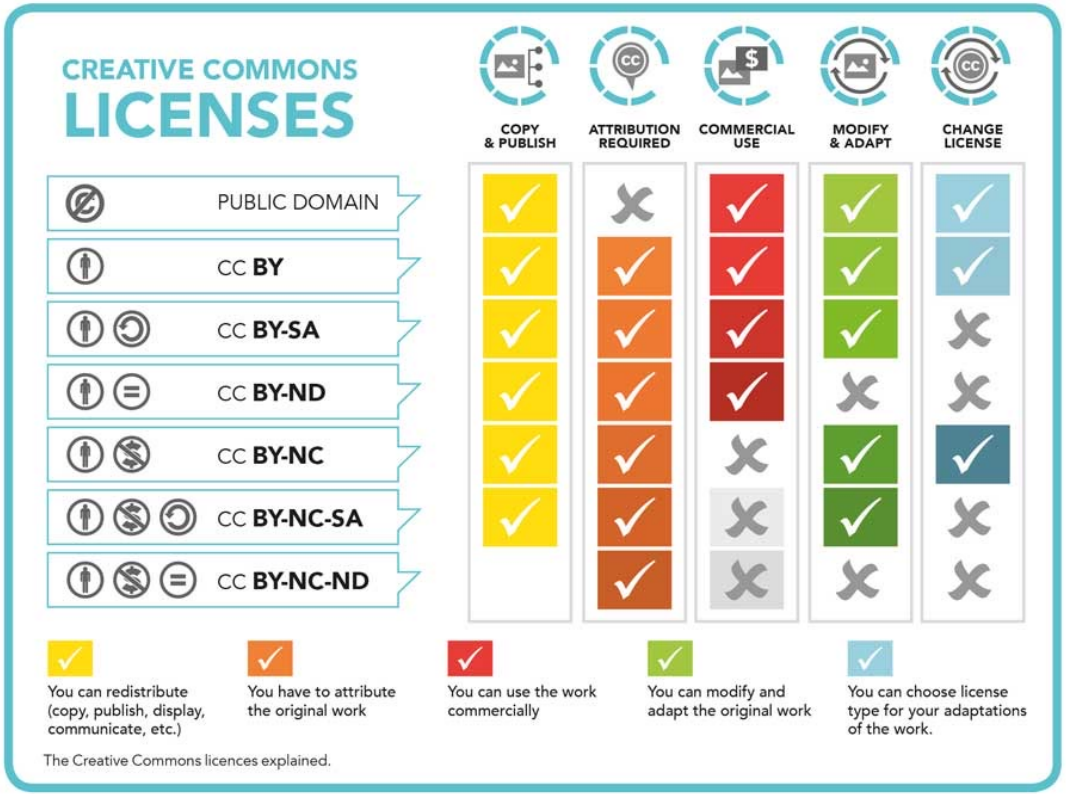


Image courtesy of [Research Outreach]

Sensitive Data and Privacy

- Remove direct identifiers only when appropriate; linkage may still identify people.
- Minimize data shared and use secure transfer and storage.
- Assess risks to individuals and groups, including inference risks.



Codebooks and Data Dictionaries

- Define variables, units, allowed values, and missing-value codes.
- Explain derived variables and transformations.
- Document study design, instruments, and collection conditions.
- Use discipline-specific terms/ controlled vocabularies consistently.

Preservation and Reuse

- Prefer stable, documented, non-proprietary formats where suitable.
- Preserve code, metadata, and explanatory documentation alongside data.
- Plan for updates, retention, and long-term stewardship.
- Make reuse conditions and provenance visible to future researchers.

Dimensions of Data Quality

- Validity: do measures represent the intended concepts?
- Completeness: which values or groups are missing?
- Consistency and accuracy depend on instruments and procedures.
- Quality is a question of **fitness for purpose** relative to a specific research goal, not to be measured by a single score.

03 | Research Infrastructure and Collaboration

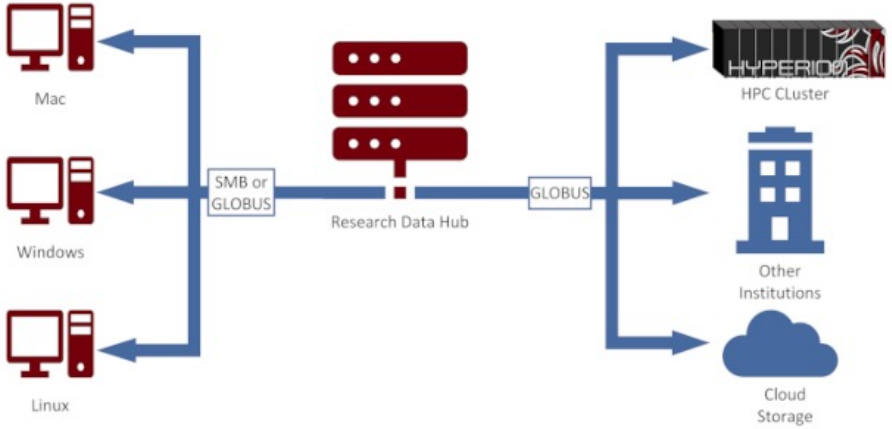
Infrastructure Across the Research Lifecycle

- Storage and backup protect active project data.
- Compute resources enable analysis and simulation.
- Repositories support discovery, sharing, and preservation.
- Persistent identifiers connect data, code, publications, and researchers.

Local to Cloud to Repository

- Choose storage based on sensitivity, scale, and access needs.
- Use secure transfer and access controls for sensitive data.
- Separate active working copies from preservation copies.
- Document where authoritative versions are maintained.

Research data infrastructure



https://sc.edu/about/offices_and_divisions/division_of_information_technology/rc/resources/research_data_storage/

Data Management Planning

- Plan data types, formats, documentation, and storage.
- Specify access, sharing, retention, and preservation arrangements.
- Assign roles, resources, and responsibilities.
- Revisit plans when the project or data risks change.

Team Roles and Data Flow

- Researchers, data stewards, librarians, and IT staff contribute different expertise.
- Define who collects, validates, documents, approves, and shares data.
- Agree on communication and change-management practices.
- Clear roles reduce gaps during handoffs.



Data Stewardship Responsibilities

- Assign owners for quality, access, security, and documentation.
- Define escalation routes for incidents or access requests.
- Track responsibilities across institutional and team boundaries.
- Stewardship continues after publication.

Sustainable Preservation and Handoffs

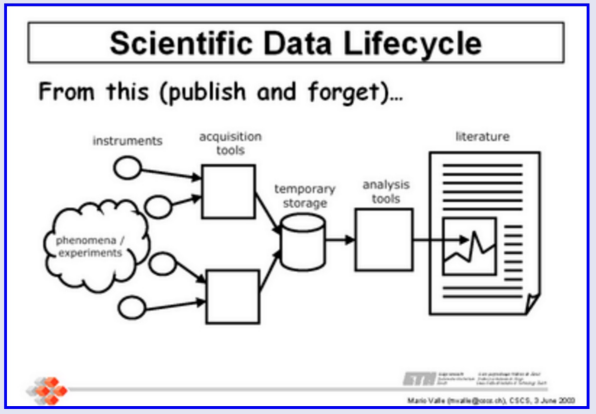
- Plan for staff turnover, project closure, and software obsolescence.
- Preserve context and documentation along with files.
- Identify long-term repository and funding responsibilities.
- Make handoffs explicit to avoid undocumented orphaned data.

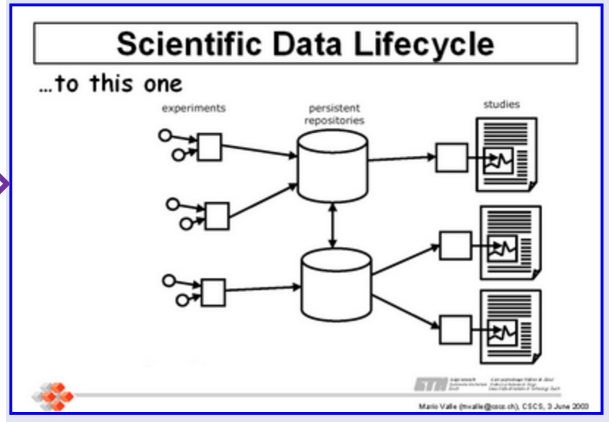
Retention, Access, and Disposal

- Set retention periods based on research, ethical, and legal needs.
- Review permissions when roles or projects change.
- Securely dispose of data when retention ends, where required.
- Maintain an auditable record of stewardship decisions.

Problems found in current scientific projects

- Limited file and directory naming schemes. Some project data repositories are simply big flat directories (**No Logical Organization**!)
- No access to important metadata in scientists' notebooks and heads.
- Un-owned data with dubious content after the end of project or PhD thesis.





Summary

- Design data collection around a clear research question.
- Choose methods that fit the question; communicate uncertainty and limits.
- Preserve provenance, documentation, and reproducible workflows.
- Share data responsibly and plan for sustainable stewardship.

Thanks!