

Research Data Management

Mara Sedlins, PhD
Data Management Specialist, CSU Libraries
mara.sedlins@colostate.edu



Overview

Data management: What and why

Documentation & Reproducibility

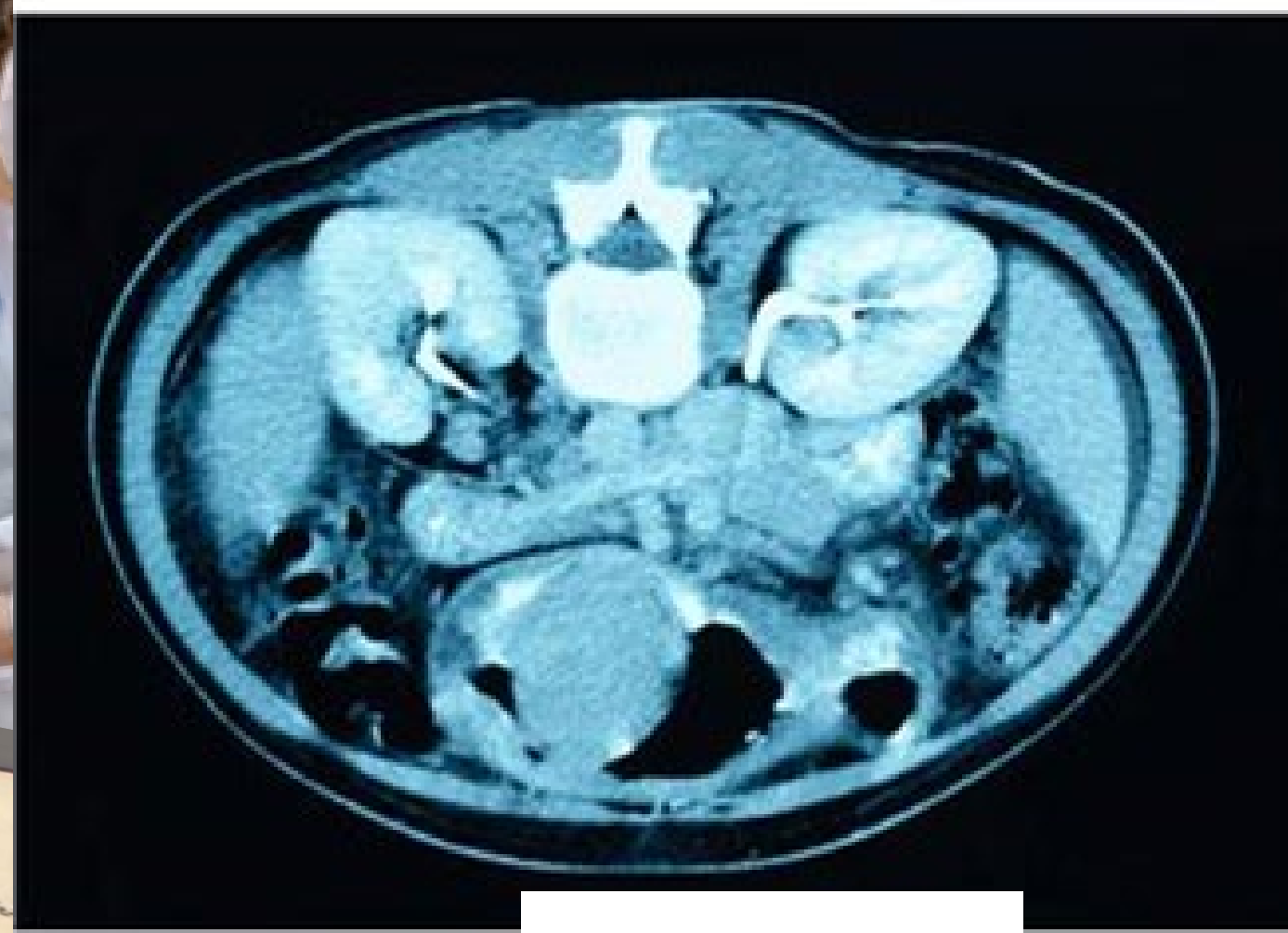
Storage & Preservation

Data Management Plan Exercise

What is data management?

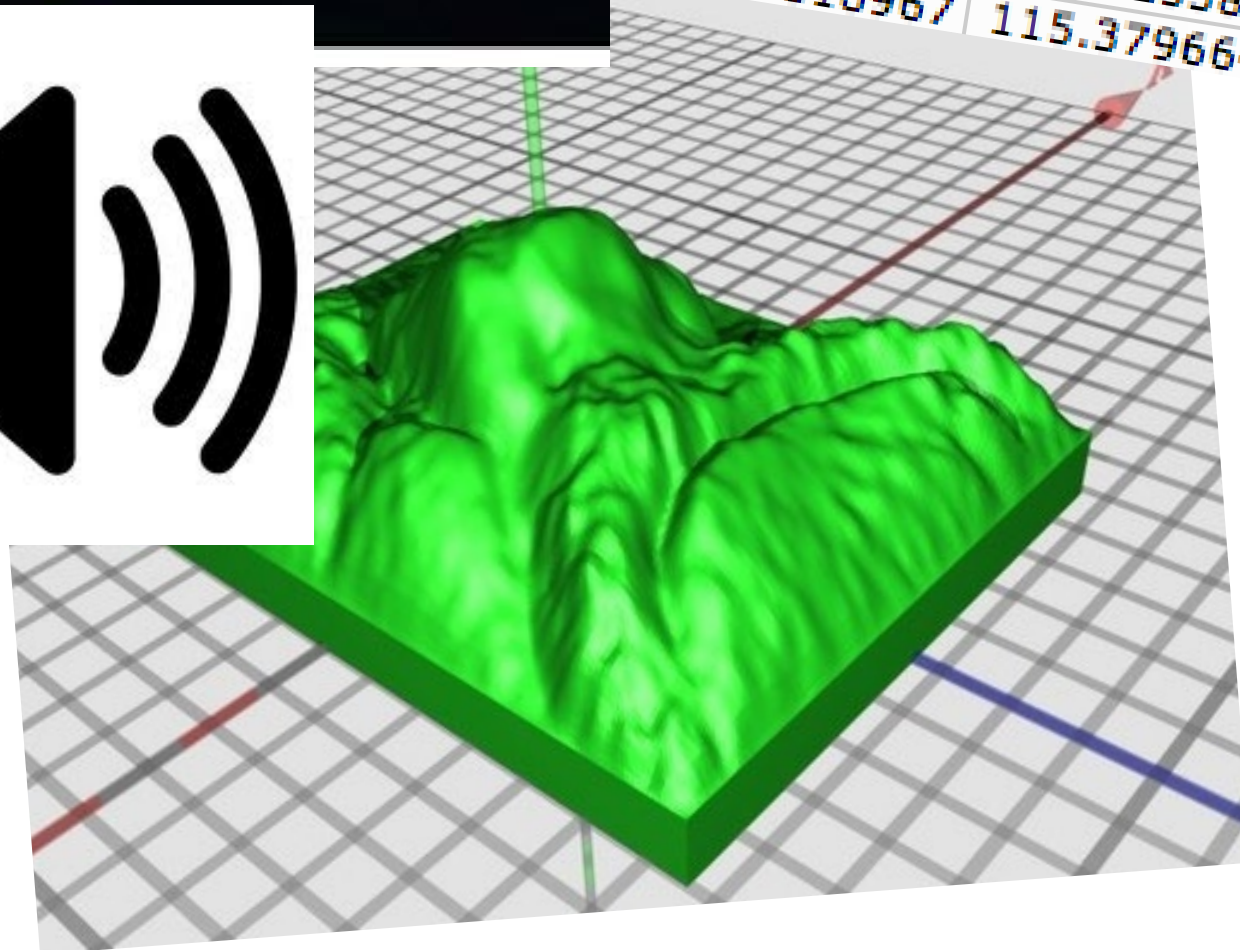
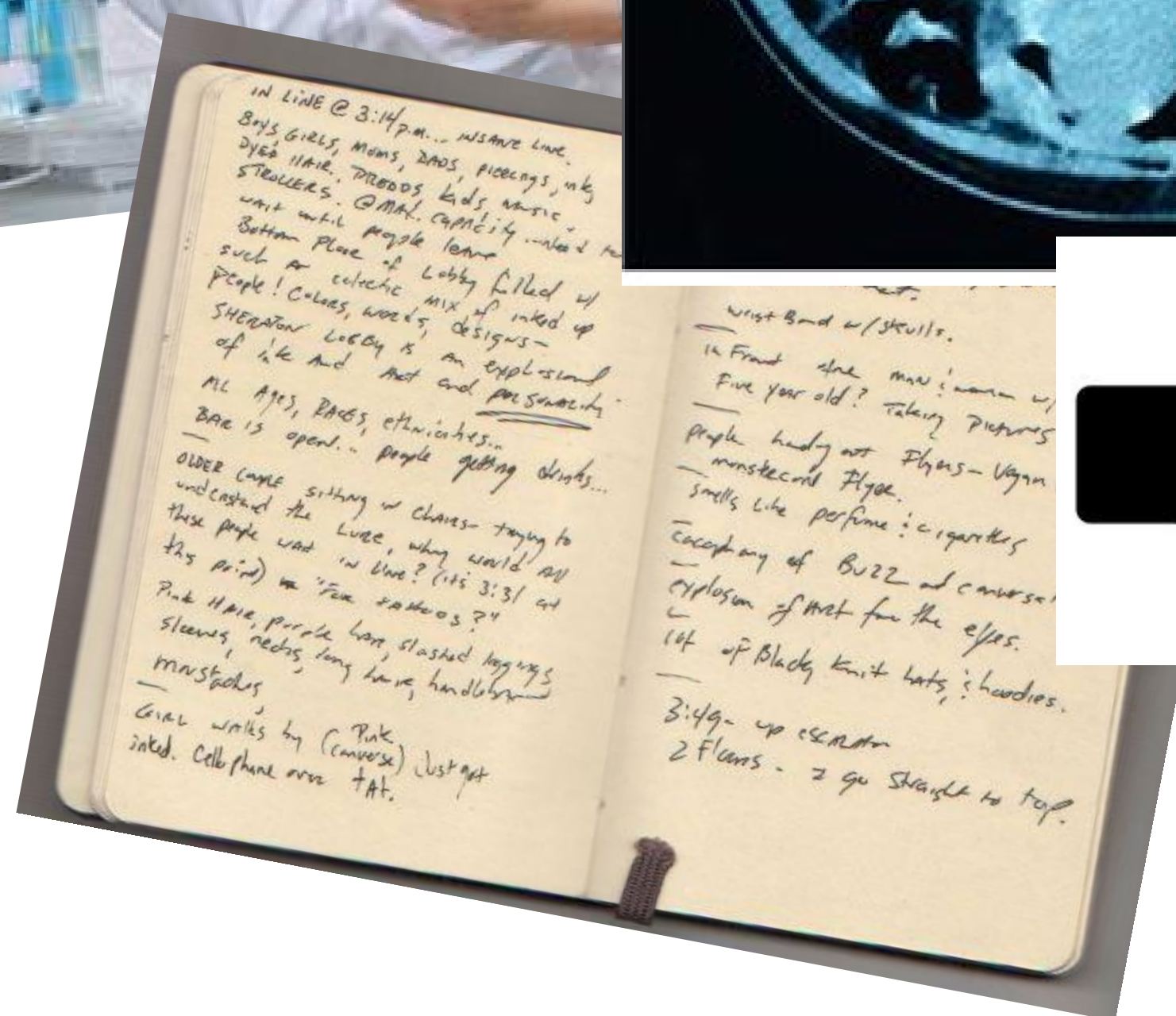
The **policies, practices and procedures** needed to manage the **storage, access and preservation** of data produced from a research project

What is data?



10/29/04

761606	129.27356	122.260995	209.6621
977679	129.534731	124.739135	176.3164
46875	135.839924	130.84732	168.2896
95502	149.510531	140.795689	120.6866
27052	140.495868	132.823819	206.1385
51598	137.880438	124.888856	189.6756
7241	131.84633	126.146789	202.4966
5374	130.691651	112.877008	140.3665
1212	121.561443	114.237637	125.2985
488	128.496503	113.302591	192.2236
813	138.880759	128.517198	108.7016
465	139.289941	129.528986	127.4065
785	135.363241	127.454638	129.6691
35	133.242253	124.704841	244.5670
17	135.159011	125.476984	169.2715
53	127.612613	124.25382	170.4015
7	122.818967	115.379664	134.9701



Research
data are
diverse!

Types of Data

- Primary / Secondary
- Qualitative / Quantitative
- Experimental / Observational

Types of Data

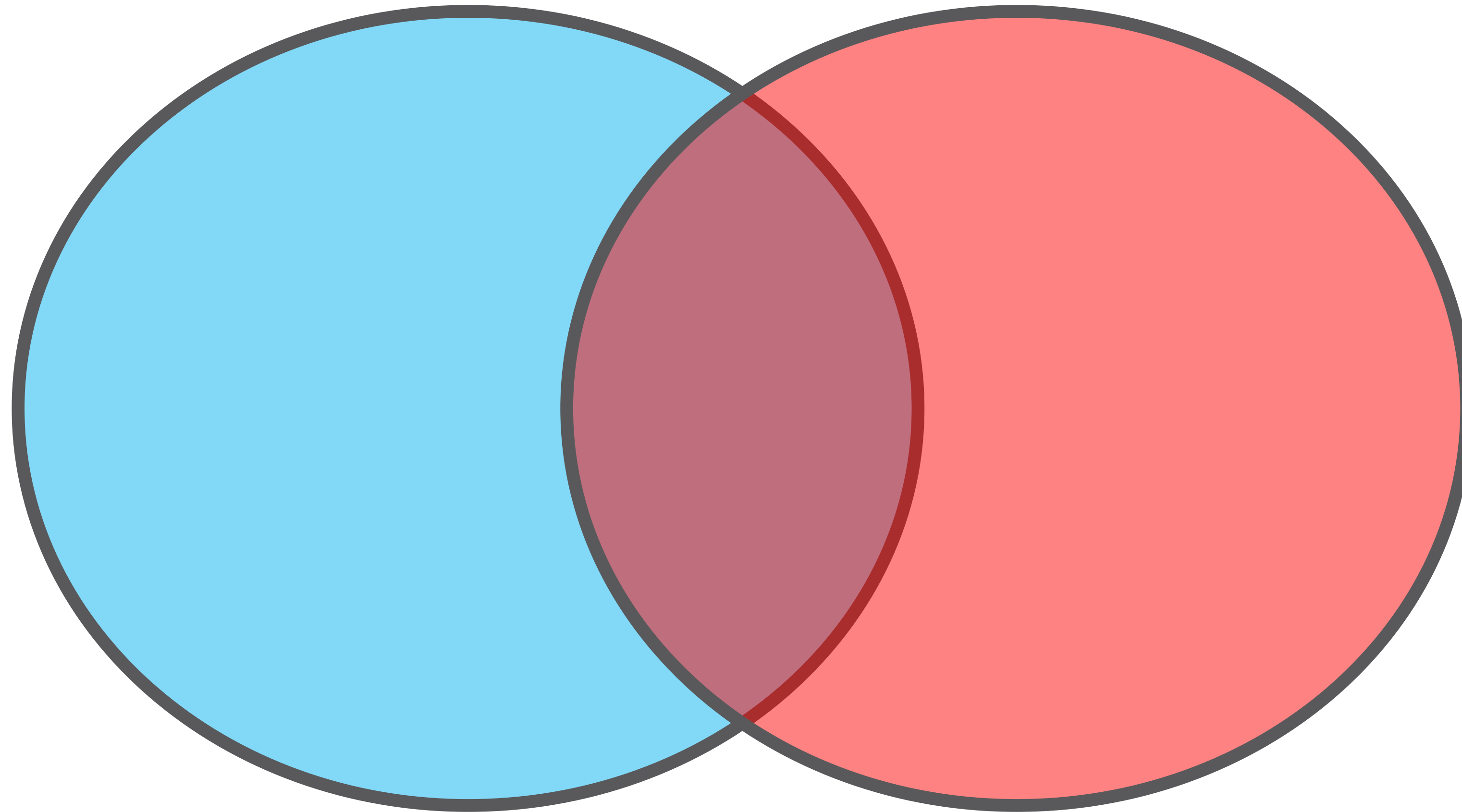
- . Instrument measurements
- . Experimental observations
- . Still images, video and audio
- . Databases
- . Quantitative data (data tables; spreadsheets)
- . Interview transcripts & text documents
- . Simulation data, models & software
- . Slides, artefacts, specimens, samples
- . Sketches, diaries, lab notebooks ...

What is data?

“The **recorded** factual material commonly accepted in the research community as necessary to **validate research findings**.”

-U.S. Office of Management and Budget (OMB) Circular A-110

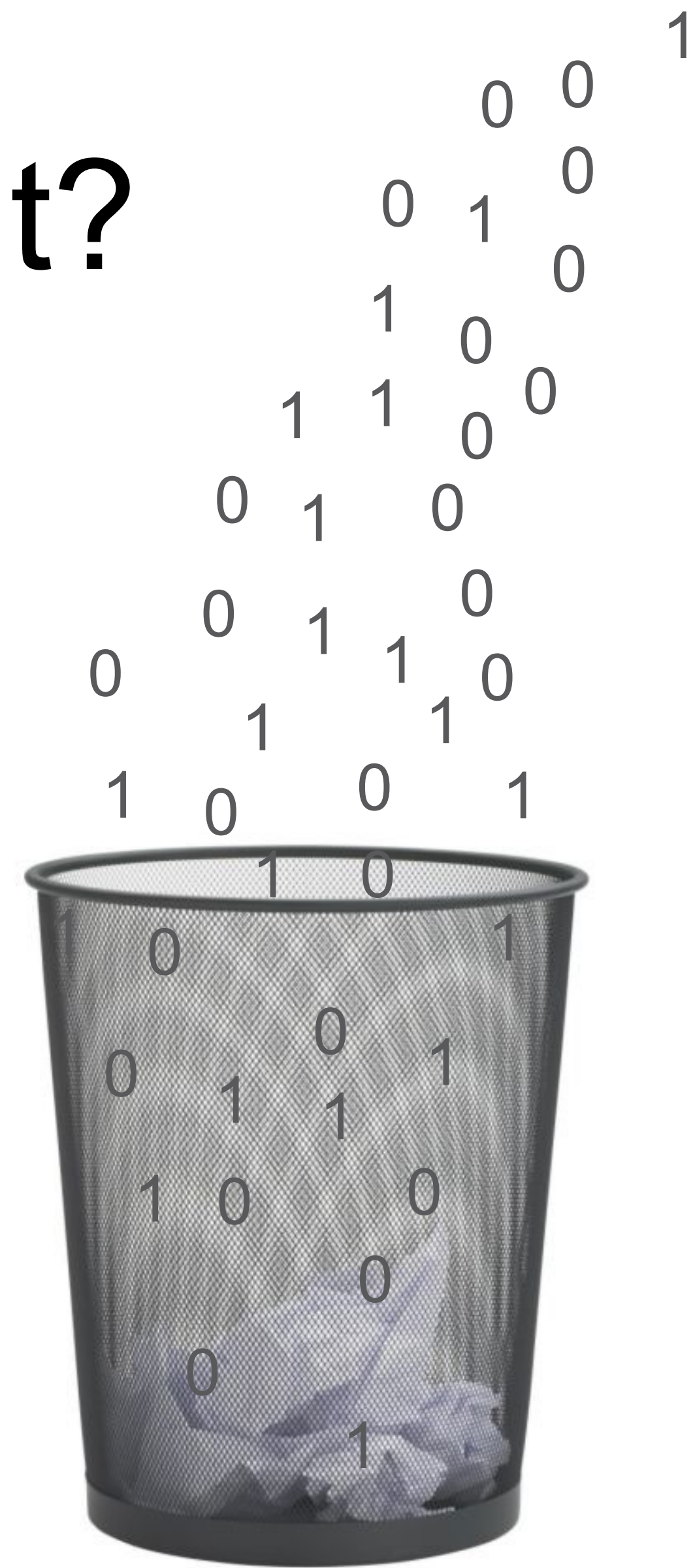
data management \neq data sharing



but the same principles apply to both

Why is data management important?

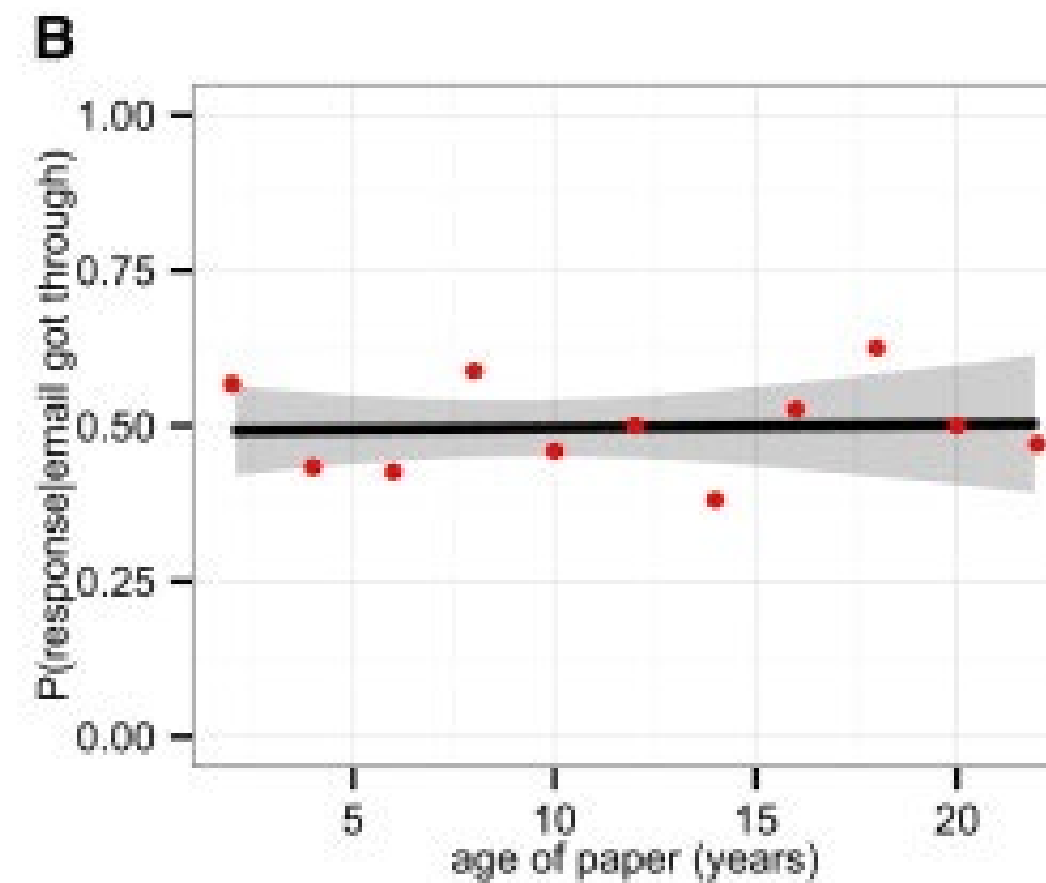
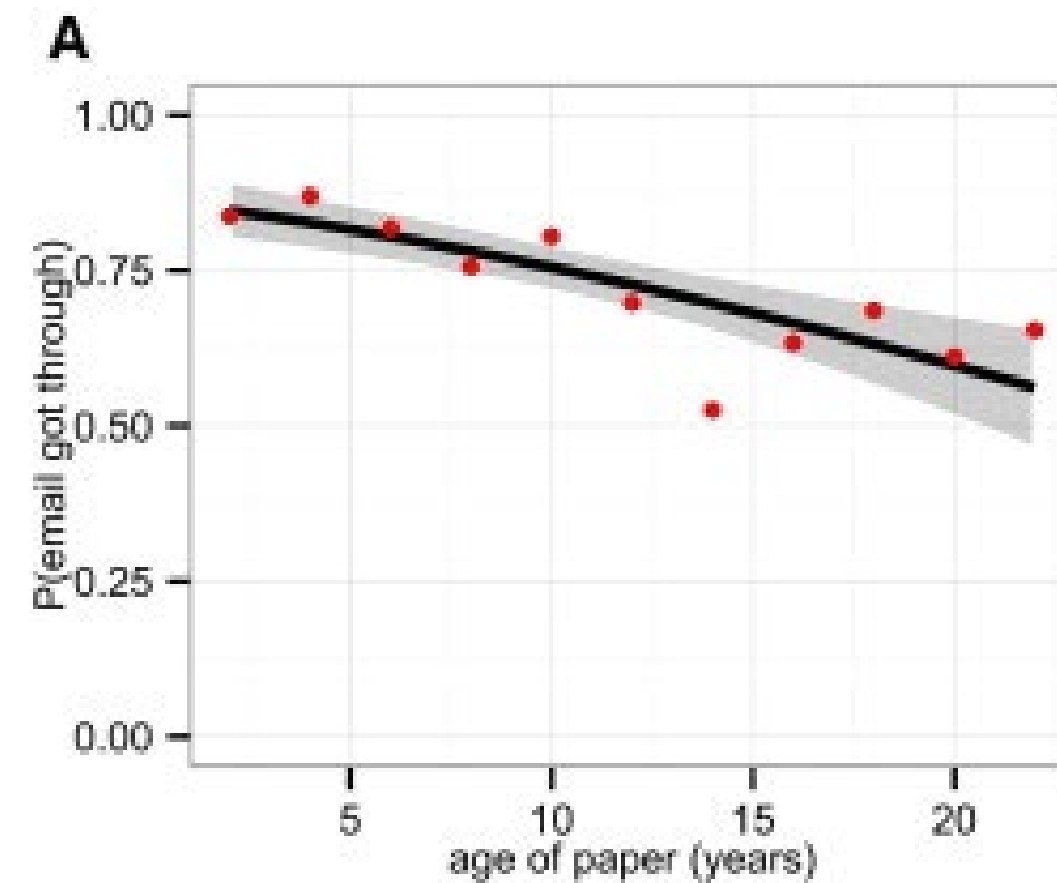
- Data management and sharing requirements
 - Funders
 - Journals
- Everything is digital → Easier to lose
- New concept → Requires training



Report

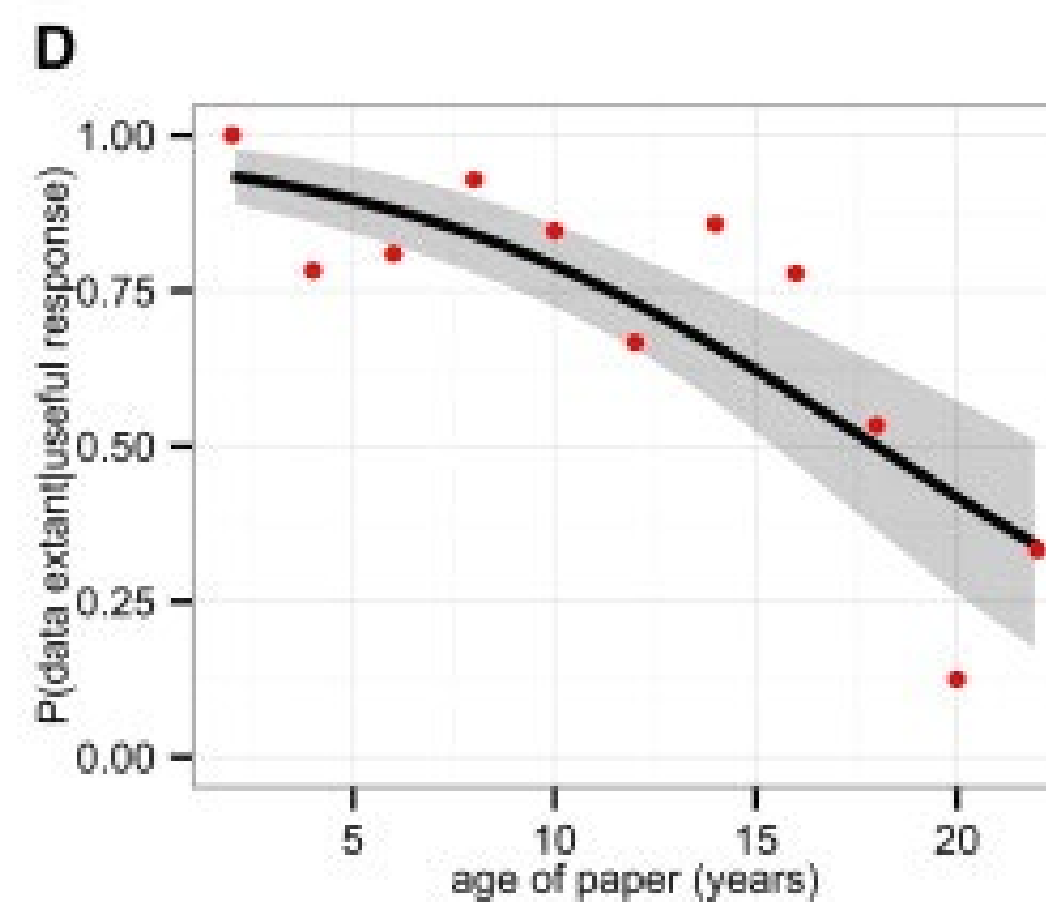
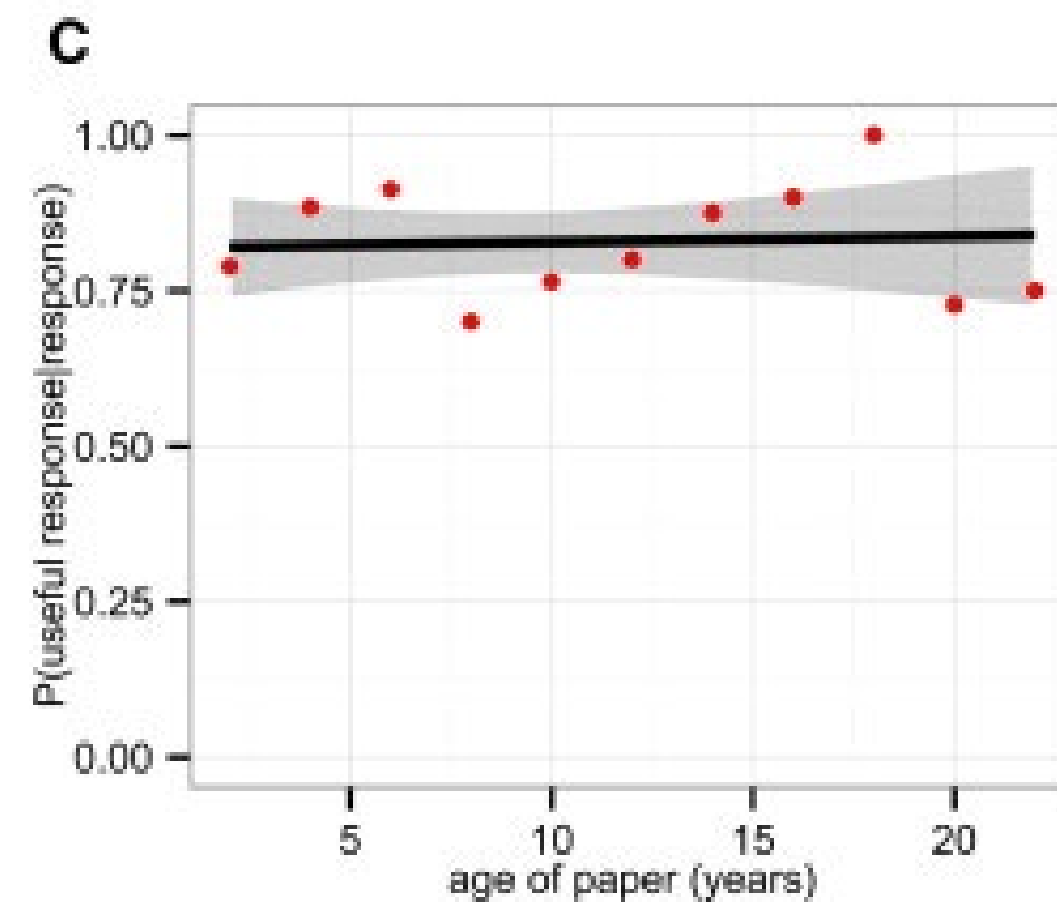
The Availability of Research Data Declines Rapidly with Article Age

Working
email



Response to
email (if
email works)

Information on
status of data



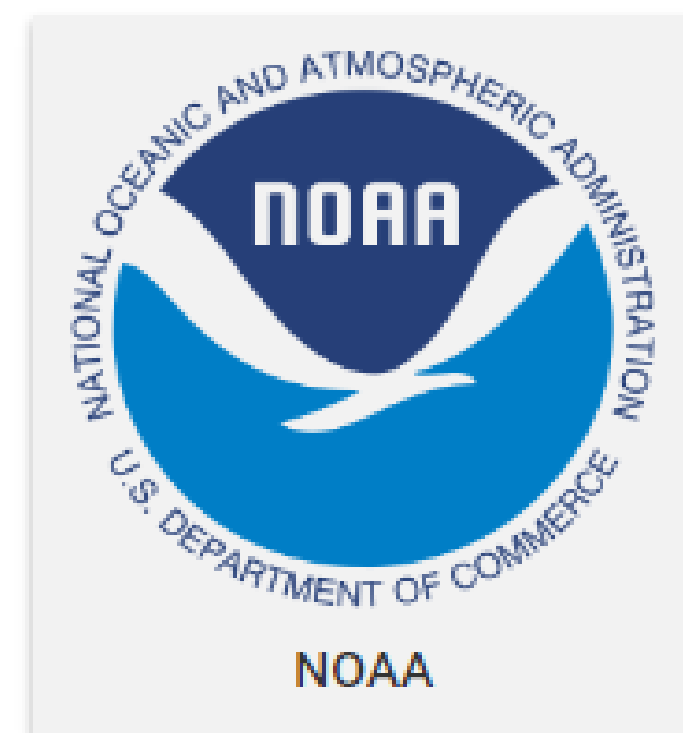
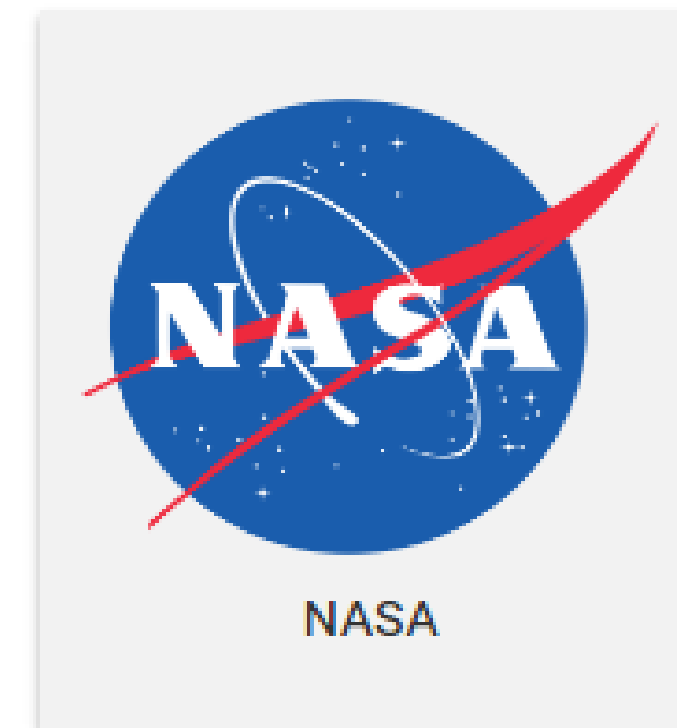
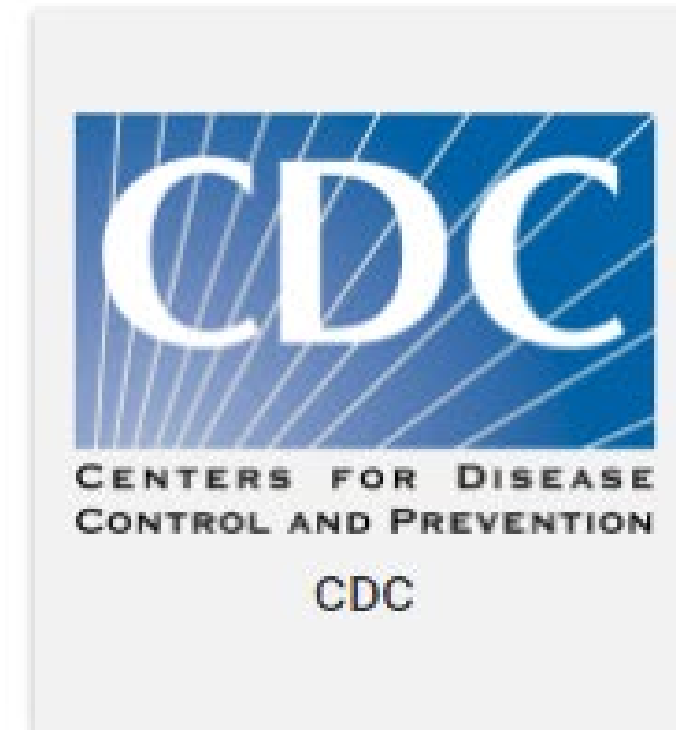
Data still exist

<https://doi.org/10.1016/j.cub.2013.11.014>

We are losing vast amounts of data



Who is responsible?



<http://datasharing.sparcopen.org/data>


2022 OSTP Memo



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

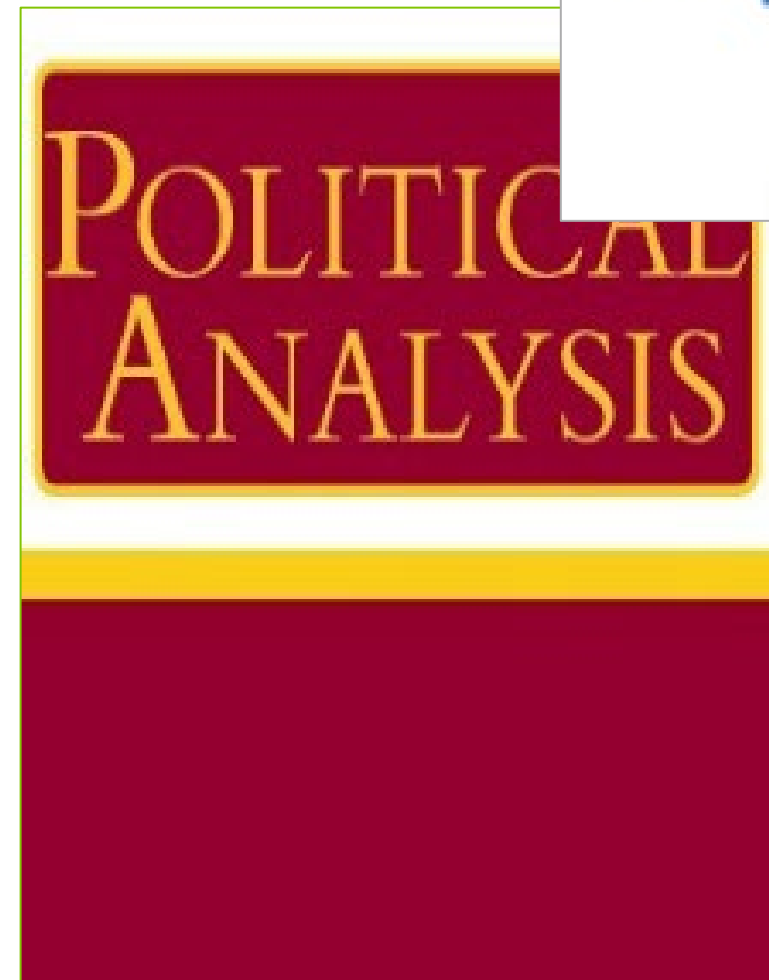
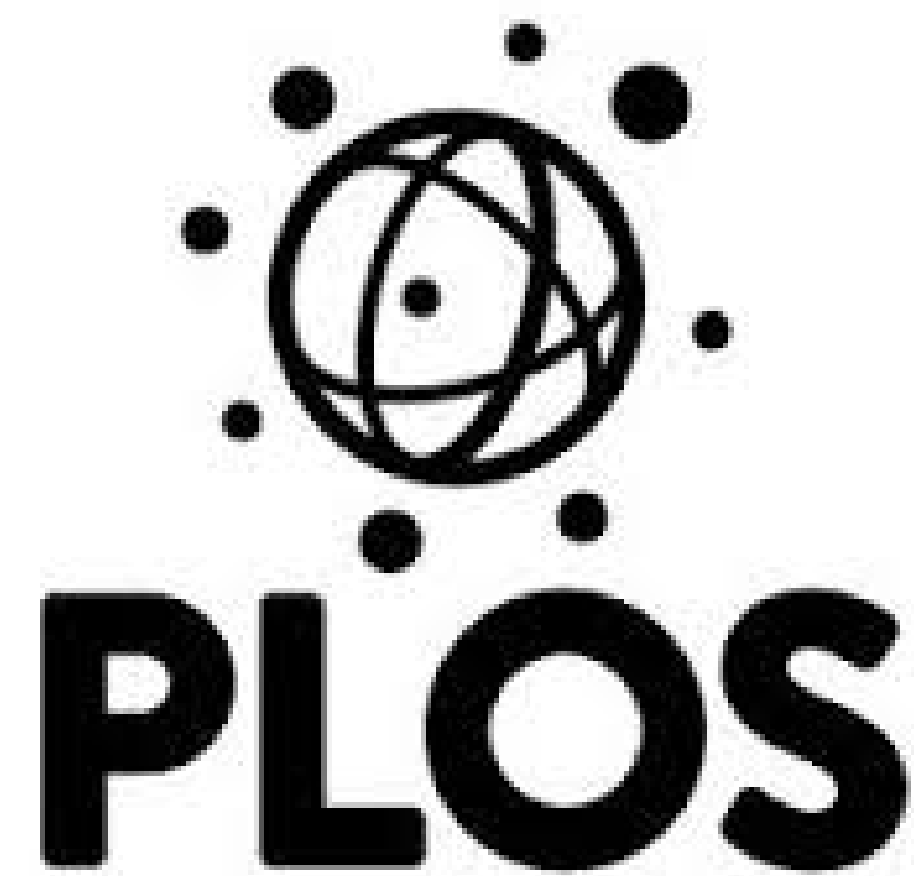
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson 
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data

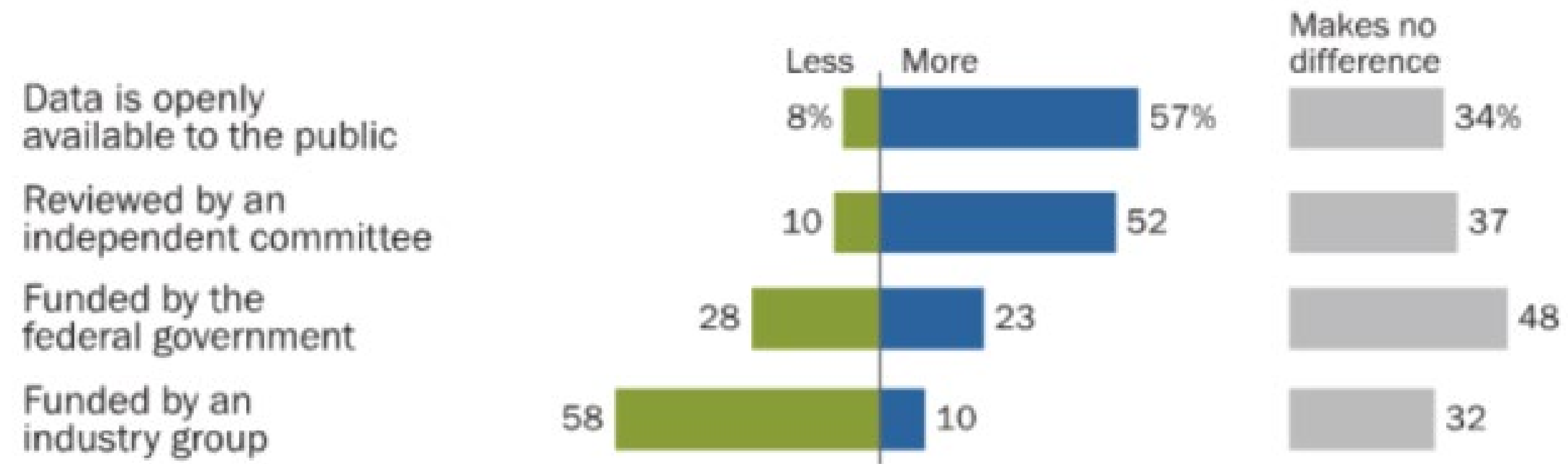


Example: <https://www.nature.com/nature-research/editorial-policies/reporting-standards>

It's good for science

- Improves research reproducibility
- Improves efficiency
- Spurs innovation
- Increases public trust in science

% of U.S. adults who say when they hear each of the following, they trust scientific research findings ...



Note: Respondents who did not give an answer are not shown.

Source: Survey conducted Jan. 7-21, 2019.

"Trust and Mistrust in Americans' Views of Scientific Experts"

PEW RESEARCH CENTER

It's good for you

- You are the future data user
- Your data get used (and cited)
- Exposure to collaborators
- More competitive grants



the life-changing
magic of tidying data

dr. tracy teal

Data Sharing and Management Snafu in 3 Short Acts:

A data management horror story

by Karen Hanson, Alisa Surkis and Karen Yacobucci.

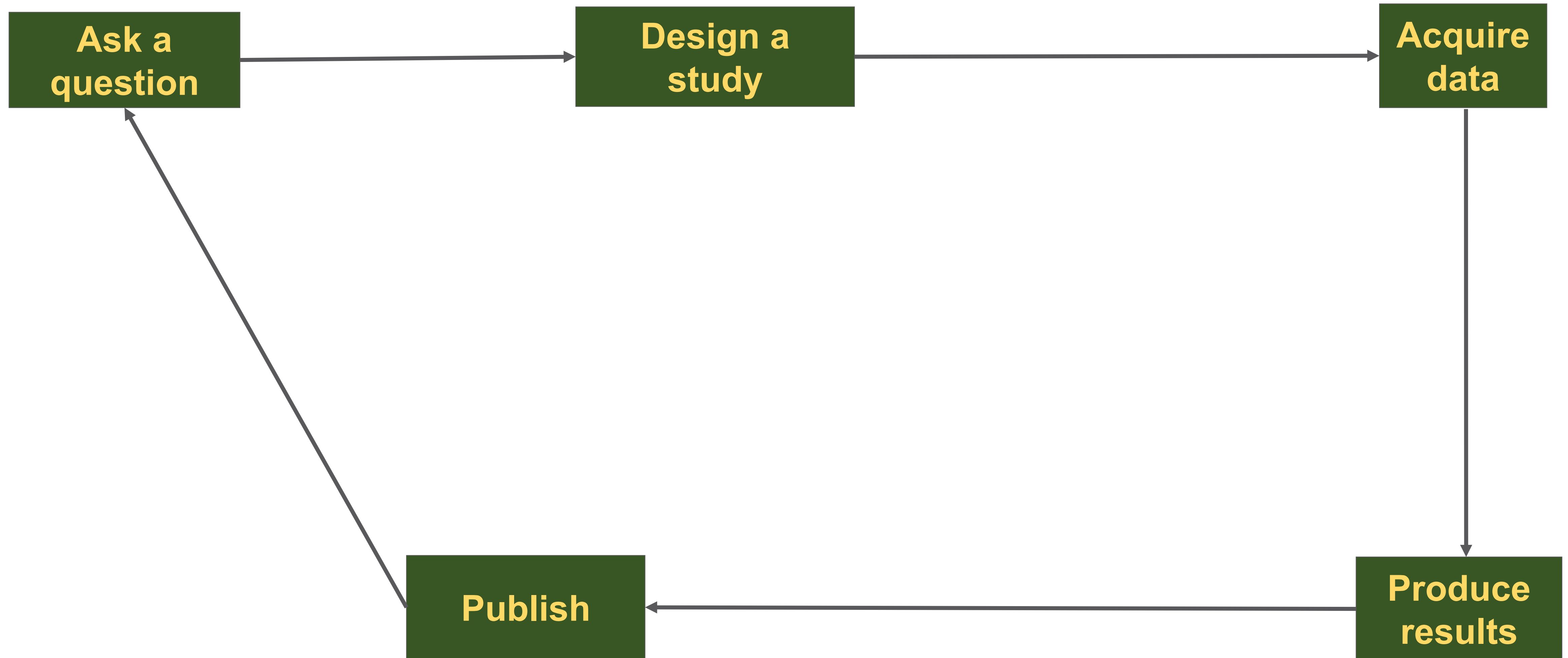
https://www.youtube.com/watch?v=66oNv_DJuPc&t=11s

Where does data
management fit into
research?

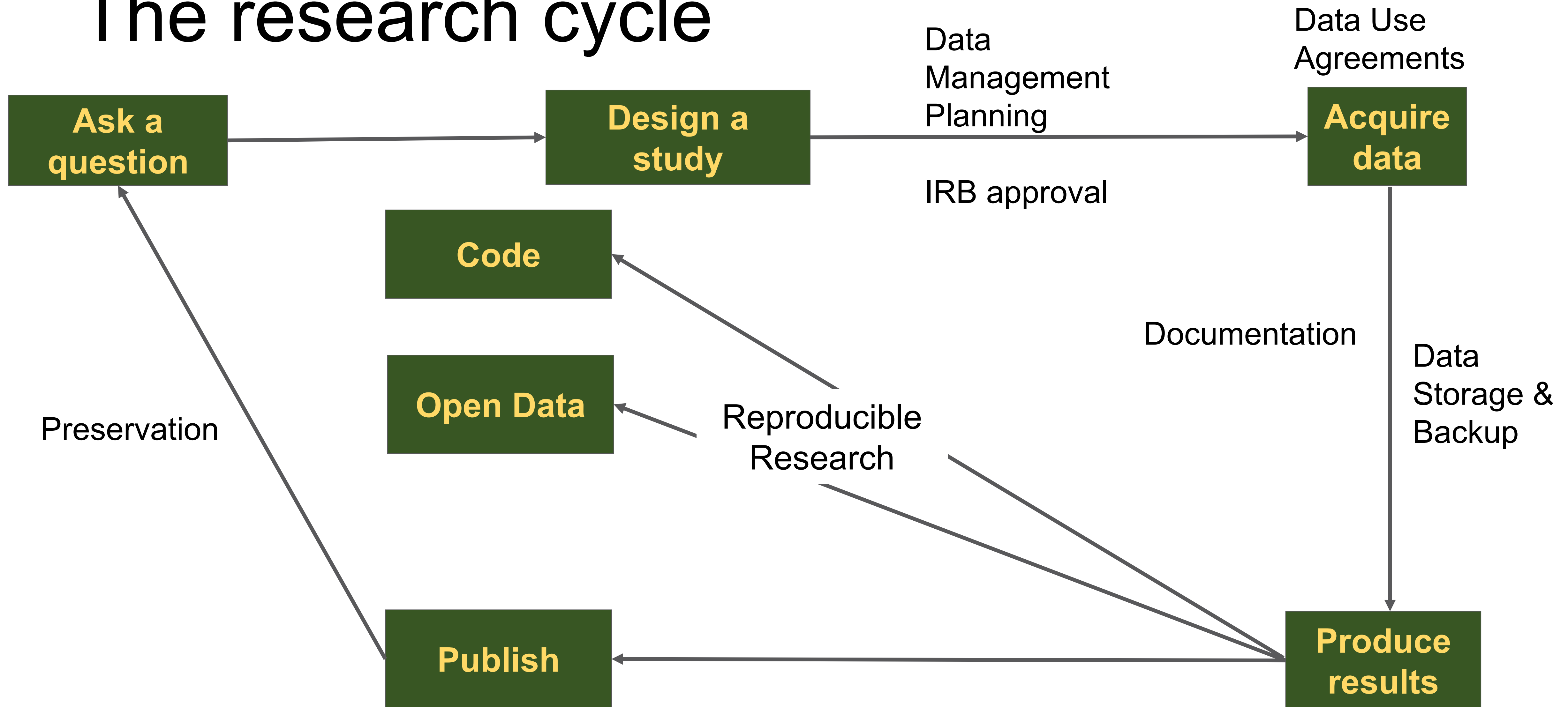
Where does data management fit into research?

Throughout the whole research cycle

The research cycle



The research cycle



The research cycle



What is a data management plan?

A description of how you plan to describe, preserve and share your research data.

Good DMPs include

- A **data inventory**, including type(s) and size
- A **strategy for describing** the data
- A plan for **ethical and legal** compliance
- A plan for **storing and securing** the data
- A method for **preservation and access** to the data
- A description of **roles and responsibilities**

Data inventory

- What **type of data** are you going to collect?
- What **file type** will be produced?
- How **stable** is the data?
- **What size** will these files be? **How many** files?
- What **other research outputs** will be produced?
 - Code/Software?
 - Templates/protocols?

Data Acquisition

- Do the data you need already exist? Do you need to collect data?
- Resources for finding existing data:
 - <https://libguides.colorado.edu/strategies/data>
 - <https://libguides.colostate.edu/statisticsources>
 - Directory of Data Repositories: re3data.org
 - Data reuse may involve an application process and/or data use agreement

File Management



Image courtesy of Flickr user [Jeffrey Beall](#)

Project directory structure

Project_1

- methods
- raw_data

- analysis

- scripts
- manuscript

- readme and/or ELN link

- Develop an informative directory structure
- Keep research materials together

Slide from: [Repro4Everyone](#)
Inspired by [‘Bioinformatic data skills’](#)
by Vincent Buffalo



Project directory structure

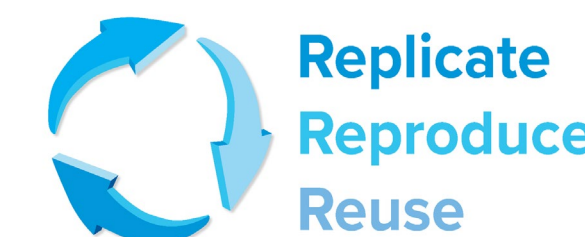
Project_1

- methods
- raw_data
 - readme
- analysis
 - analysis_method_1
 - 2017
 - 2018
 - analysis_method_2
- scripts
- manuscript
 - text
 - version_1
- readme and/or ELN link

Specific content in each category for Project #1

Raw data,
Data analysis, and
Manuscript

Slide from: [Repro4Everyone](#)
Inspired by [‘Bioinformatic data skills’](#)
by Vincent Buffalo



Project directory structure

Project_1

- methods
- raw_data
 - readme
- analysis
 - analysis_method_1
 - 2017
 - 2018
 - analysis_method_2
- scripts
- manuscript
 - text
 - version_1
- readme and/or ELN link



- Always keep raw data!

- Always backup data

(3x and synchronized: 3 unique locations - cloud, server, personal drive)

Slide from: [Repro4Everyone](#)
Inspired by ['Bioinformatic data skills'](#)
by Vincent Buffalo

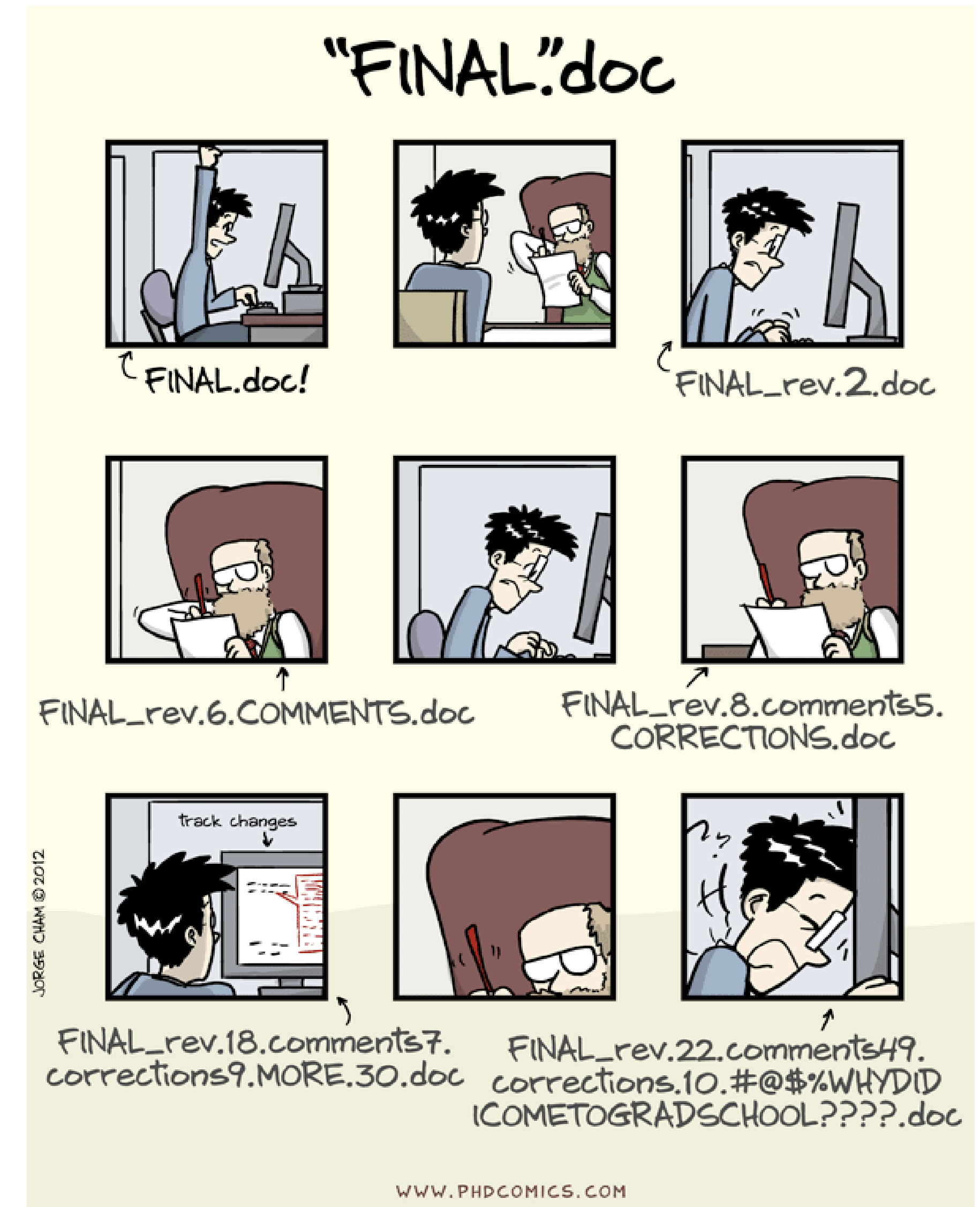
File naming conventions

- What did you call the last file you generated?
- Did you have rules?

<https://guides.lib.purdue.edu/c.php?g=353013&p=2378293>

<http://kbroman.org/dataorg/>

Slide from: [Repro4Everyone](#)



File naming conventions

The rules don't matter; that you have rules matters

- Include date in yyyy-mm-dd format
- Use meaningful abbreviations
- Have group identifiers
- Document your decisions
- Be consistent
- Use version numbers

guides.lib.purdue.edu/c.php?g=353013&p=2378293, kbroman.org/dataorg/

Slide from: [Repro4Everyone](#)

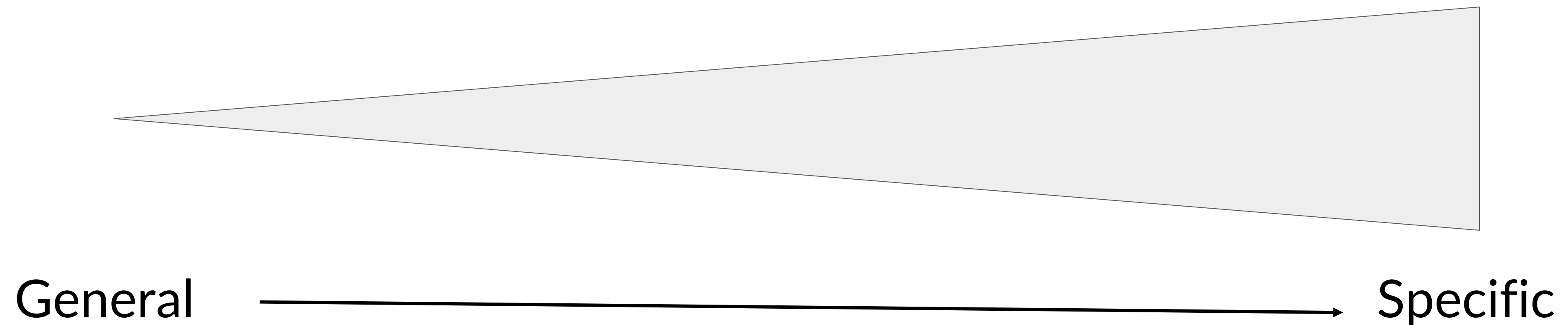


File naming conventions

- Example

20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx

Date Project Experiment Type ID Version



<https://guides.lib.purdue.edu/c.php?g=353013&p=2378293>

Slide from: [Repro4Everyone](#)

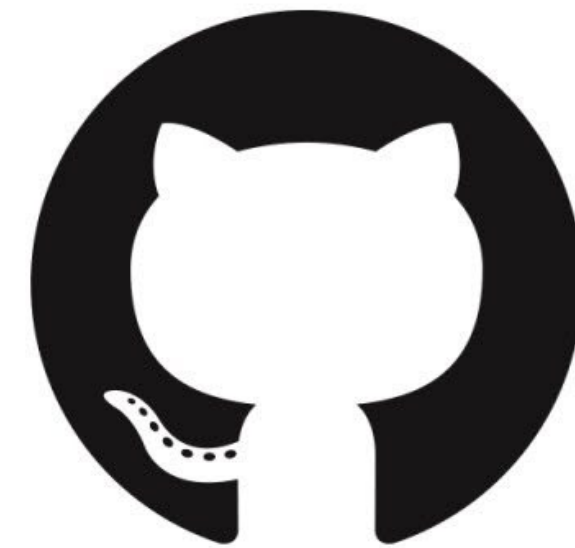


Version control – code & software



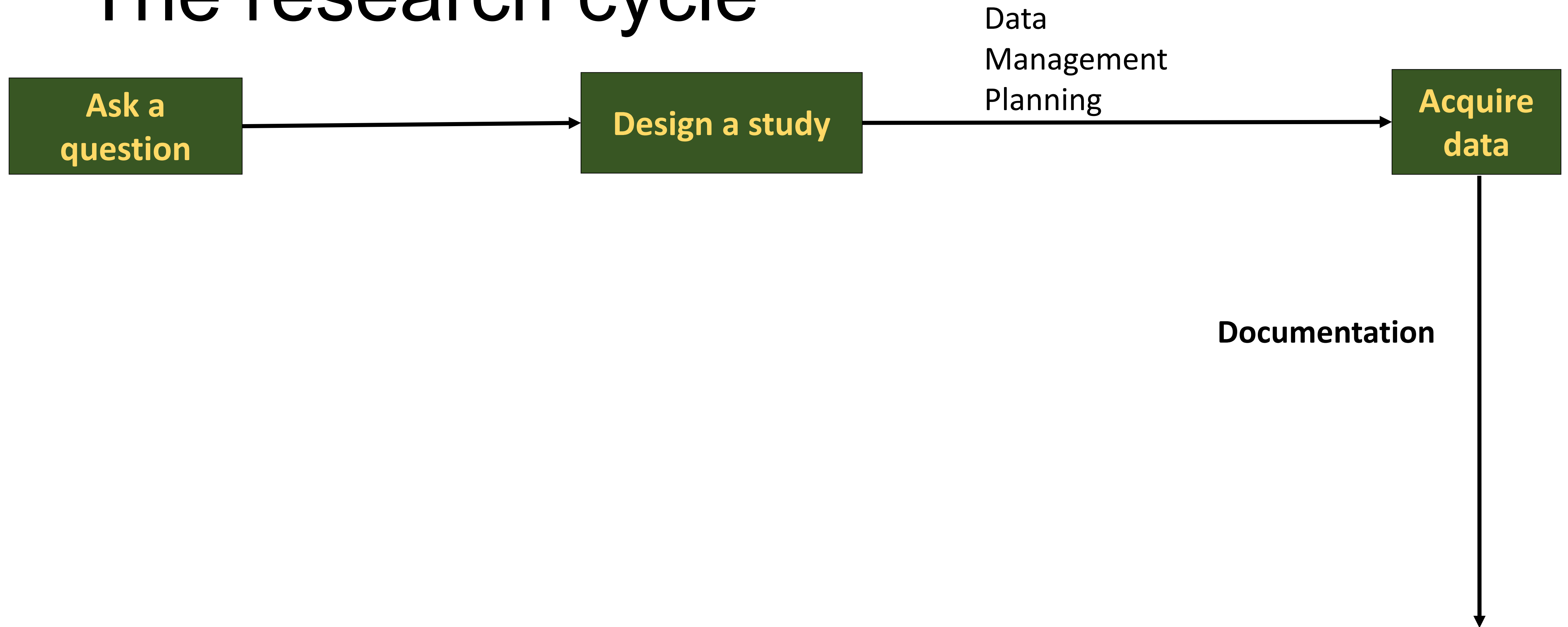
git

+



GitHub

The research cycle



A strategy for describing the data

- **Metadata:** Relevant information for re-creation and re-use
- Can be as simple as a text file



README files

- Where the data came from
- Data inventory
- Describe the **contents** of data files
- Template:
<https://data.research.cornell.edu/content/readme>

This DATSETNAMereadme.txt file was generated on YYYYMMDD by NAME

----- GENERAL INFORMATION -----

Title of Dataset:

Author Information (Name, Institution, Address, Email)

Principal Investigator:
Associate or Co-investigator:
Alternate Contact(s):

Date of data collection (single date, range, approximate date): <suggested format YYYYMMDD>

Geographic location of data collection: <City, State, County, Country and/or GPS Coordinates or bounding boxes>

Information about funding sources or sponsorship that supported the collection of the data:

----- SHARING/ACCESS INFORMATION -----

Licenses/restrictions placed on the data, or limitations of reuse:

Recommended citation for the data:

Citation for and links to publications that cite or use the data:

Links to other publicly accessible locations of the data:

Links/relationships to ancillary or related data sets:

----- DATA & FILE OVERVIEW -----

File list (filenames, directory structure (for zipped files) and brief description of all data files):

Relationship between files, if important for context:

Additional related data collected that was not included in the current data package:

If data was derived from another source, list source:

If there are there multiple versions of the dataset, list the file updated, when and why update was made:

----- METHODOLOGICAL INFORMATION -----

README files

DO

- Keep it concise, yet informative
- Use headings, line breaks, tables, and bullet points for readability
- Define the variable list, including full names and definitions of column headings for tabular data, units of measurement, explain abbreviations and any empty cells
- Describe any scripts, code, notebooks and the software used to run them (e.g., R, Python, Mathematica, MatLab) as well as the software versions, including packages, that you used to run those files
- Provide links to publications that cite or use the data, other publicly accessible locations of the data and/or the related research article
- List other sources, if any, that the data was derived from

DON'T

- Assume that variables, abbreviations/shorthand, acronyms, units, scoring keys, etc. are always used in the same way or universally understood
- Include author names or other identifying information (initials, email addresses, ORCIDs) if the journal follows a double-blind review process
- Include the Abstract or Methods sections of your manuscript as a substitute for explaining your data
- Include statements that are phrased in a way to suggest any legal imperative for attribution (e.g., “required,” “must”) or other conditions for reuse; instead *encourage* potential users to contact you or cite the data for additional information or potential for collaboration

<https://blog.datadryad.org/2023/10/18/for-authors-creating-a-readme-for-rapid-data-publication/>

Codebooks

- Define the **variables** and their **units**
- Explain the **formats** for dates, time, geographic coordinates
- Define any **coded values** and **missing values**

	A	B	C	D	E
	#	Variable name	Type	Label	Values
	1	SurveyID	Numeric	Internally generated in Access, where data stored	None
	2	SurveyRefNumber	Numeric	Number assigned to households on actual survey (pdf) ranges 1-706	None
	3	AIMAG_CODE	String	Aimag or provinces code	{11, Arkhangai}...
	4	AIMAG_NAME	String	Aimag or province name where household located	None
	5	SOUM_CODE	String	Soum or district code	{1106, Ikh-Tamir}...
	6	SOUM_NAME	String	Soum or district name where household located	None
	7	Bag	String	Bag or sub-district name or its number	None
	8	ORG_CODE	String	Code assigned to User Group with aimag, soum codes contained	None
	9	ORG_NAME	String	Organization name to which household affiliated	None
				CRRM organizational status: formal vs	

Formal Metadata Schemas

- Many **discipline specific** metadata standards
 - DDI: <https://ddialliance.org>
 - ISO 19110/19139 for geospatial data: <https://www.fgdc.gov/metadata/iso-standards>
 - EML: <https://eml.ecoinformatics.org>
- Search for **other standards**:
 - <http://www.dcc.ac.uk/resources/metadata-standards>
 - <https://fairsharing.org/standards/>

```
<?xml version="1.0"?>
<eml:eml
  packageId="eml.1.1" system="http://knb.ecoinformatics.org"
  xmlns:eml="eml://ecoinformatics.org/eml-2.0.0">

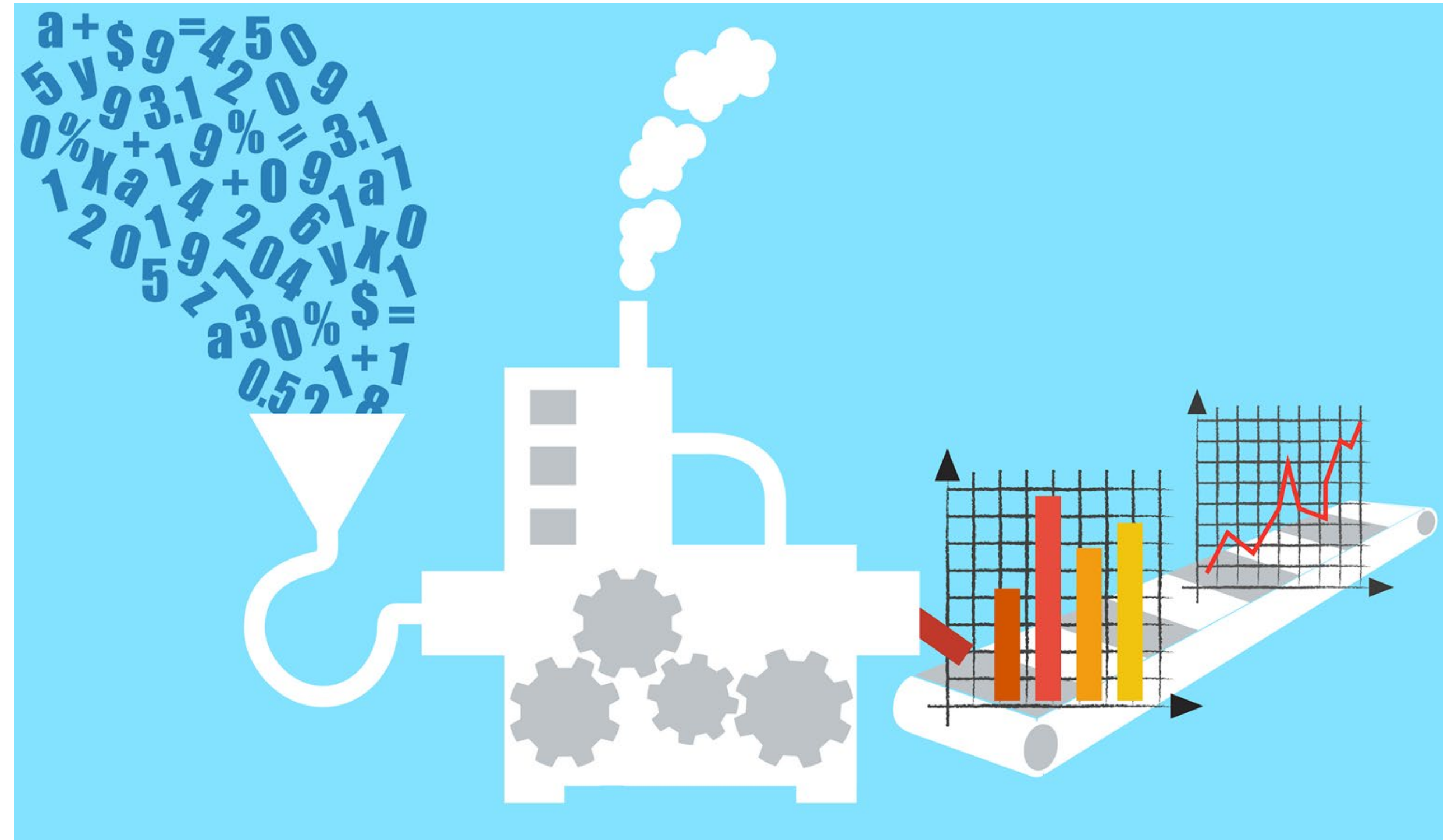
  <dataset>
    <title>Biodiversity surveys for Lesser Tree Frogs at
      Barro Colorado Island (BCI) from 1994 to 1999</title>
    <creator id="23445" scope="document">
      <individualName>
        <givenName>Jane</givenName>
        <surName>Smith</surName>
      </individualName>
      <electronicMailAddress>jane@data.org</electronicMailAddress>
    </creator>
    <contact>
      <references>23445</references>
    </contact>
  </dataset>
</eml:eml>
```


Document your analysis

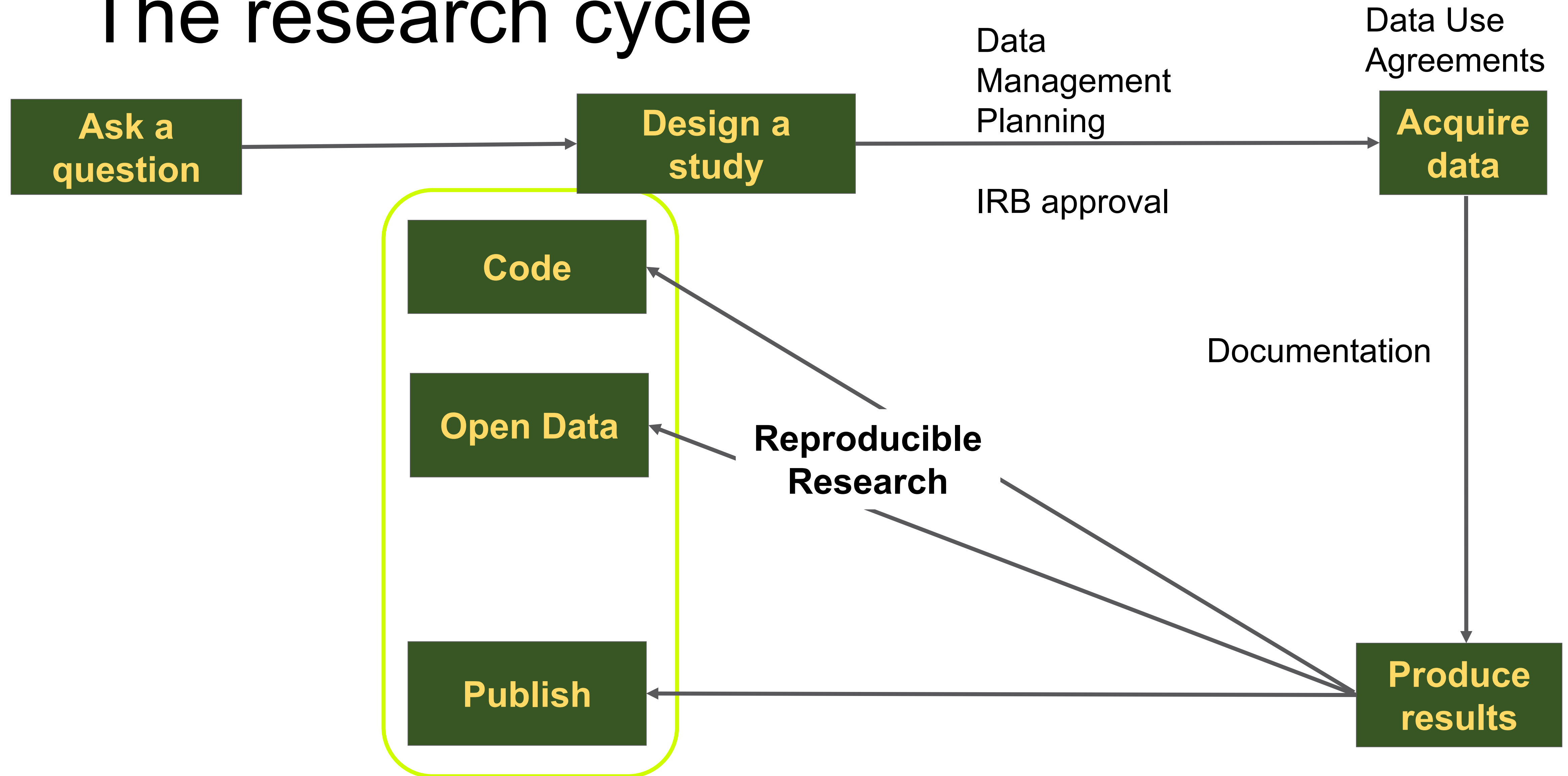
- KEEP AN ORIGINAL “RAW” DATA FILE
- Record everything you do with your data
- Can be included in the README file or as an extra spreadsheet tab

Reproducible research

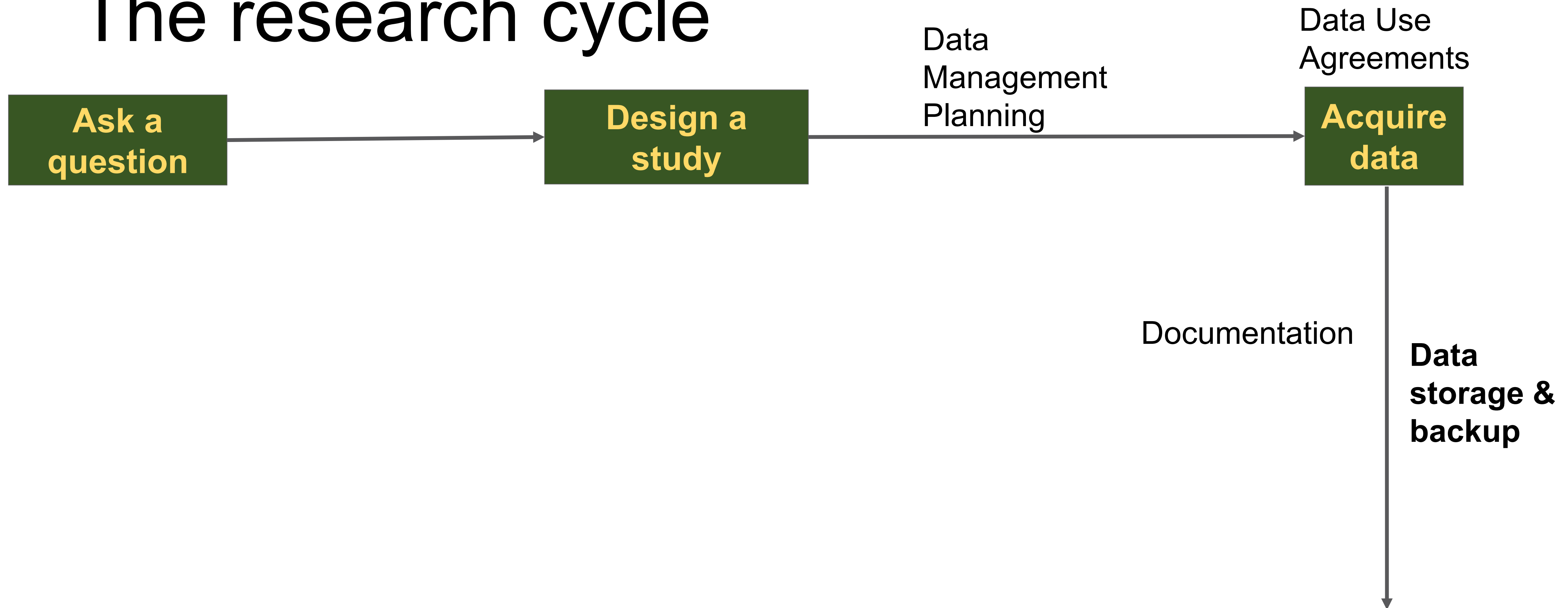
- Automate your analysis in computer code, e.g.:
 - R
 - Python
 - SPSS syntax
- Assists with repetitive tasks



The research cycle



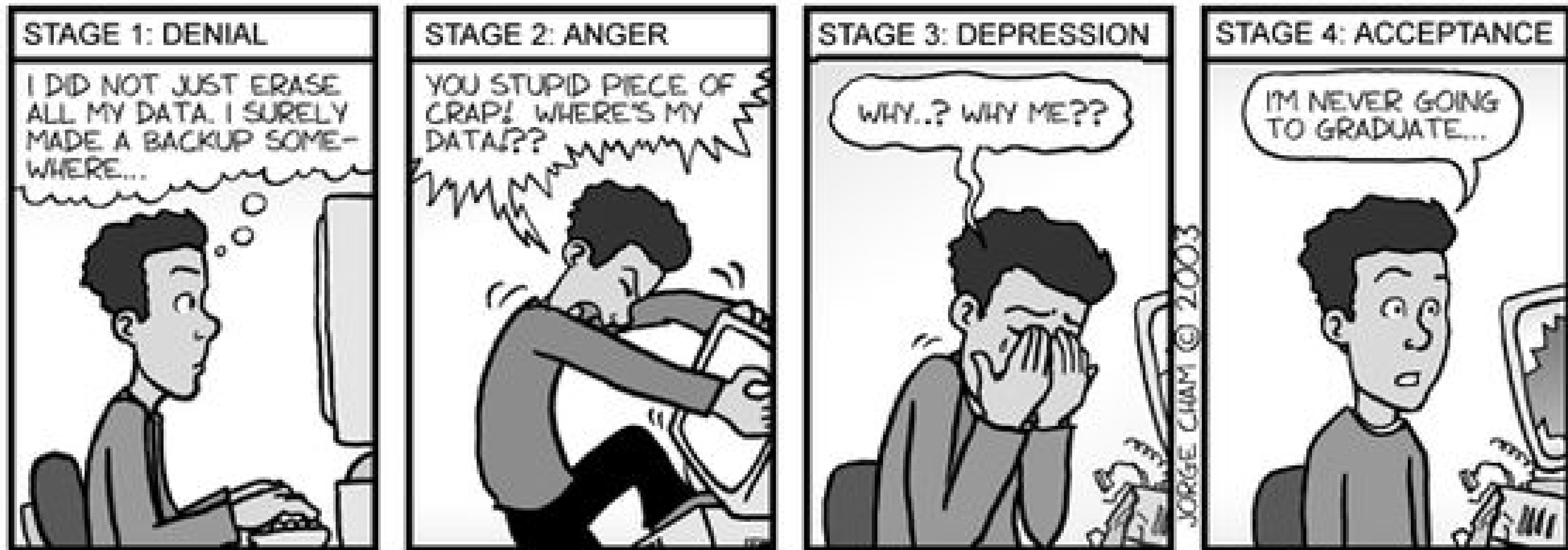
The research cycle



Backup

THE FOUR STAGES OF DATA LOSS

DEALING WITH ACCIDENTAL DELETION OF MONTHS OF
HARD-EARNED DATA



www.phdcomics.com

Data storage & backup

Short Term

- . During the project
- . Frequent changes
- . Easily accessible*

Long term

- . After the project is over
- . Little to no changes
- . Can be “put away”

*to authorized people

Data storage & backup

Short Term

- . During the project
- . Frequent changes
- . Easily accessible*

Long term

- . After the project is over
- . Little to no changes
- . Can be “put away”

*to authorized people

How sensitive is your data?

- What your data are determine where you can put it.
- Different types of data are regulated and have specific security requirements
 - Ex: PII, PHI, student data, human subjects data
- IRB protocols or data use agreements can dictate security requirements

Backup recommendations

3-2-1 Rule:

Three copies

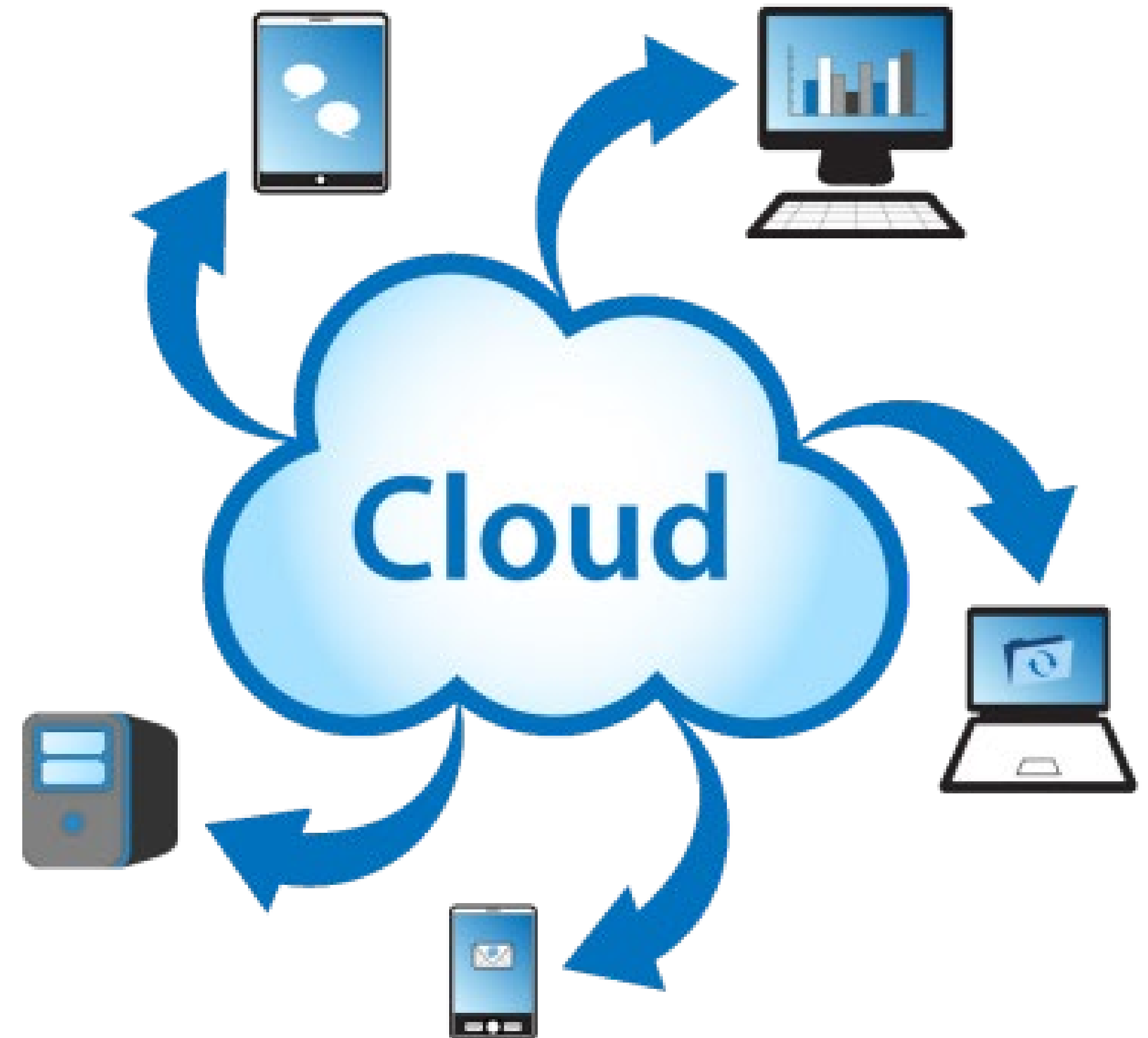
- One primary and two backups

Two formats/media

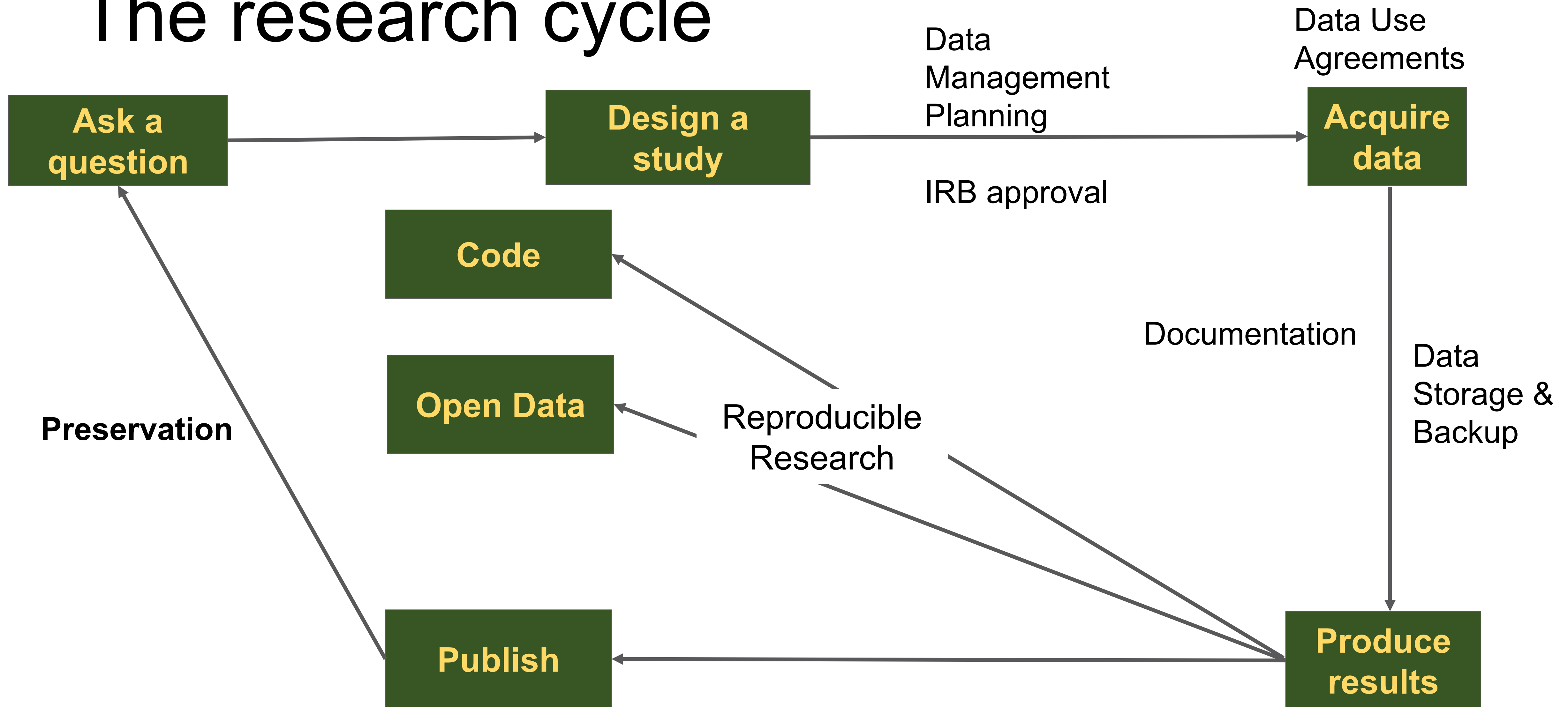
- E.g. external hard drive & cloud storage

One off site

- Where is your cloud storage located?



The research cycle



Preservation questions

- What will you store?
- Who will be in charge?
- How long will you store it?
- Where will you store it?
 - Multiple copies



Photo by [tookapic](#)

Data storage & backup

Short Term

- . During the project
- . Frequent changes
- . Easily accessible*

Long term

- . After the project is over
- . Little to no changes
- . Can be “put away”

*to authorized people

Preservation at project end

- Put away your toys!
- Include **documentation**
- Save in **archival formats**
- Link data to other research output



Image from [pixabay](https://pixabay.com/)

Archival data formats

- Avoid proprietary formats
 - Example: Excel/SPSS files
- Use common data standards in your field
- Example: .csv for tabular data



Sharing data in Repositories

Where should your data go?

- Funder specified? No?
- Journal specified? No?
- Discipline specified? No?
- Institutional repository with data?
- General repository

Sharing data in Repositories

- **Discipline specific**

- <http://re3data.org/>
- <http://FAIRsharing.org>



- **General**

- Dryad - <https://datadryad.org/>
- Zenodo - <https://zenodo.org/>
- [Comparison chart](#)

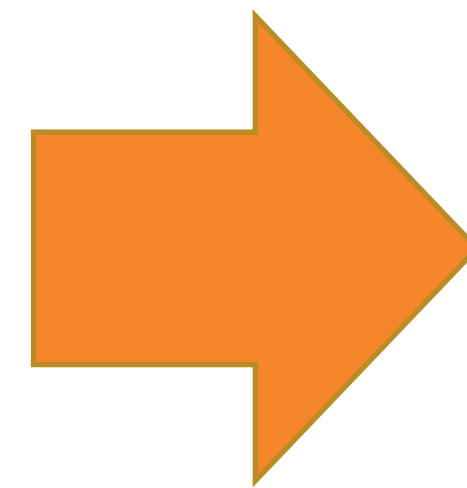


- **Institutional Repository**

- CU Scholar - <https://scholar.colorado.edu/>



Sharing research code and software



<https://guides.github.com/activities/citable-code/>

DMP Exercise

- Go to dmptool.org
- Log in or create a free account
- If you are a CU Boulder affiliate:
 - Log in using your CU Boulder email address to access CU-specific guidance
- Go to "Create Plan"
- Select a DMP Template relevant to your field (if you're having trouble choosing, check "No funder associated ...")
- Look through the prompts and attempt to answer them for your own research

DMP Exercise

- Were any of the prompts confusing or difficult to answer?
- What questions do you have? Is there anything you would like clarification on?

Data Management Tools

- [DMPTool.org](https://dmp-tool.org/)
 - Funder templates, general guidance
- [Open Science Framework](https://osf.io/)
 - Free web app for project management and sharing scholarly output
- [Git](https://git-scm.com/) and [Github](https://github.com/)
 - Version control: collaborate and track changes in code
- [OpenRefine](https://openrefine.org/)
 - Free tool for cleaning data
- File renamers (e.g. [BulkRenameUtility](https://www.bulkrenameutility.co.uk/))
 - Rename many files with just a few clicks

Help and Consultations

- One-on-one or small group consults
- Review draft DMPs and README files
- Help navigate data policies
- Find data repositories
- Advice on file formats, etc.
- CU Boulder:
 - <https://www.colorado.edu/libraries/research-assistance/data-services>
- CSU:
 - <https://lib.colostate.edu/services/data-management/>