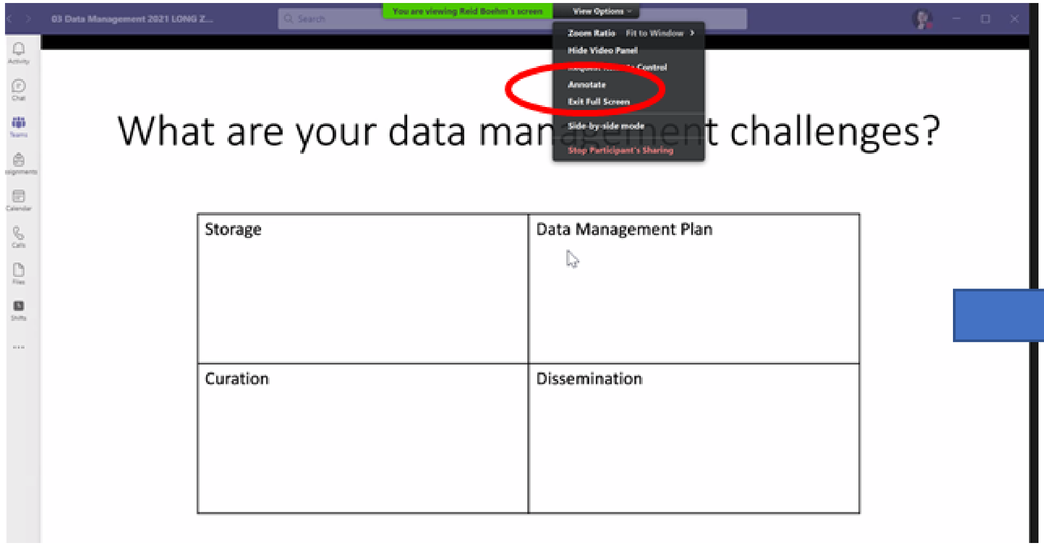
Data Management

Claudia Neuhauser, PhD, Division of Research

Santi Thompson, MA, MLIS, University Libraries

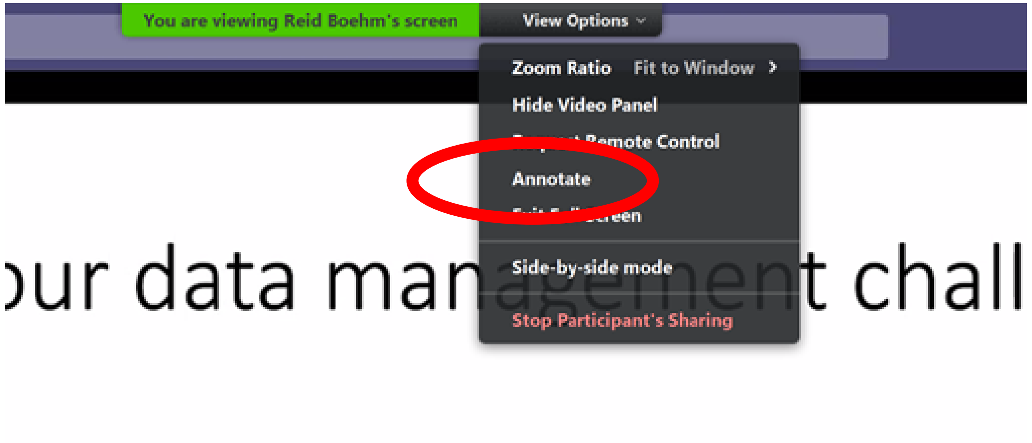
Fall 2023

Audience Interactions

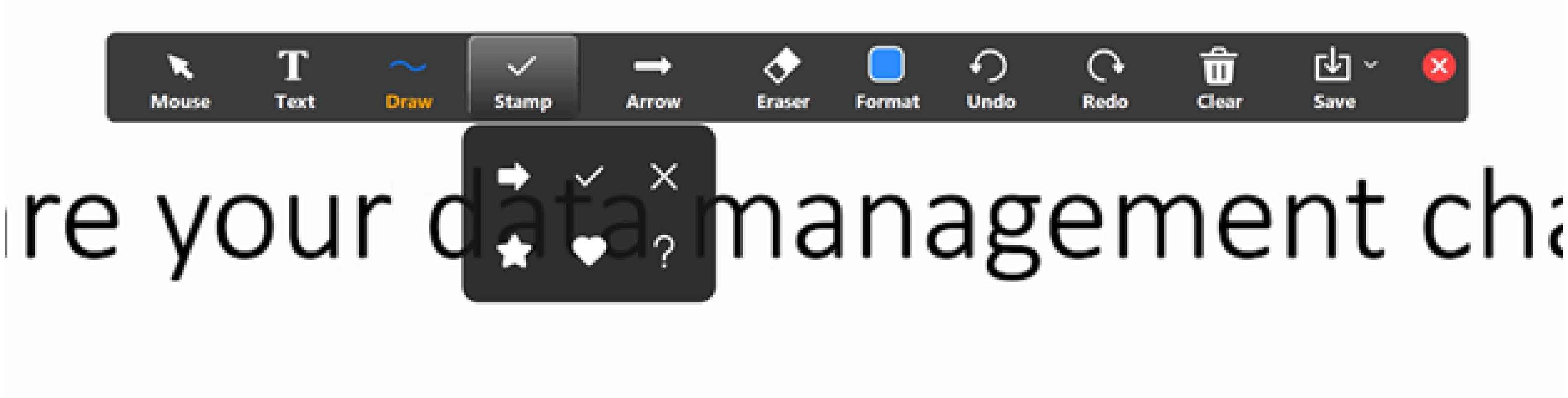


What are your data management challenges?

Storage	Data Management Plan
Curation	Dissemination



our data management chall



re your data management ch

Section 1

Agenda for the Session (Santi)

What are your data management challenges?

Storage	Data Management Plan
Curation/Archiving	Dissemination

Check which of the following are considered research data

Email communication with colleagues	
Personal notes	
Laboratory samples	
Preliminary analysis	
Computer code used for analysis	
Raw files from instruments	

Research Data

- **Recorded factual material** commonly accepted in the scientific or scholarly community as **necessary to validate research findings**, excluding preliminary analyses, drafts of scholarly or scientific work, plans for future research, peer reviews, communications with colleagues and physical objects (e.g., laboratory samples).

Email communication with colleagues	
Personal notes	
Laboratory samples	
Preliminary analysis	
Computer code used for analysis	
Raw files from instruments	

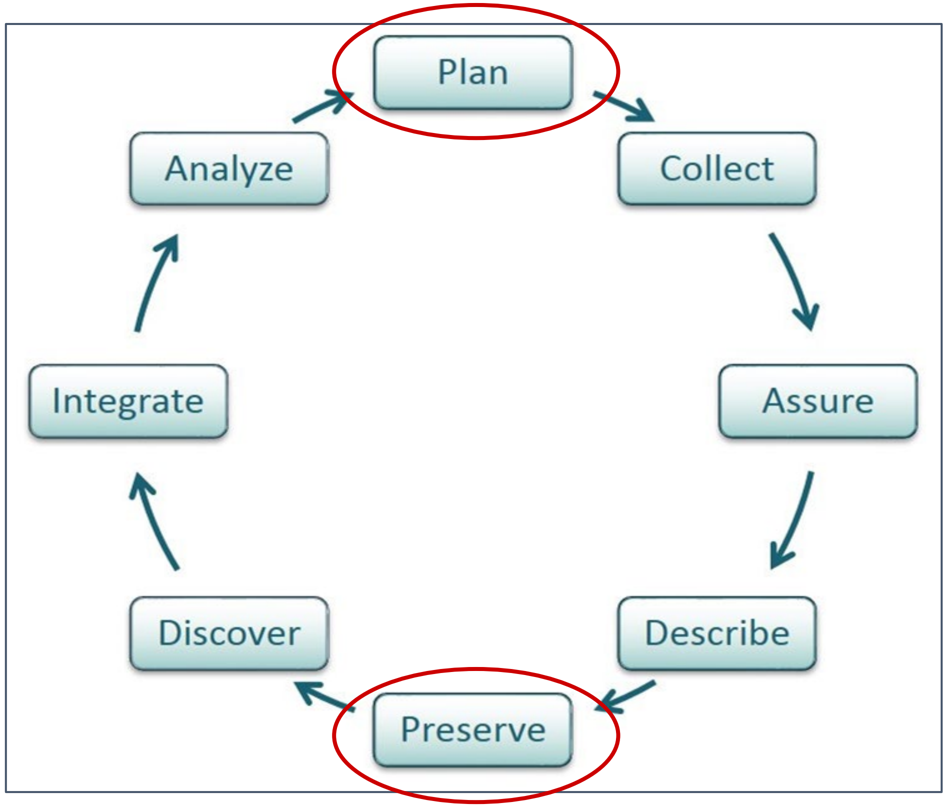
Why should you plan to manage your data?

Why plan?

- Funder requirements
- Institutional policy
- Maximize integrity in collaborations within and across institutions
- Mitigate error and loss
- Avoiding unforeseen costs
- Be able to return to the data after years of not having looked at the data
- Getting a handle on the complexity of data prior to collecting data

Components of managing research data

Planning at the outset eases the management of research outputs long term.



Section 2

Data Corruption and Poor Data Management (Claudia)

Have you lost any data previously?

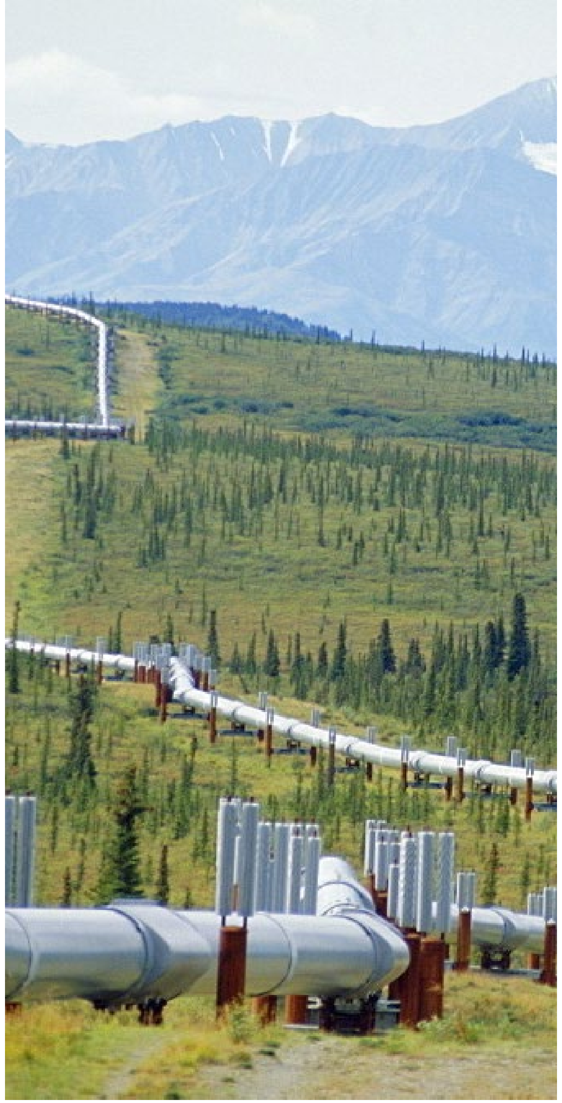
YES	NO

Data Loss: Disaster

- The Library of Alexandria in Egypt had the most complete collection of ancient literature and scholarly work when it burned down about 2,000 years ago. We lost thousands of scrolls containing advanced and irreplaceable works of mathematics, astronomy, physics, poetry etc. The data loss was so great we're still hearing about it today.

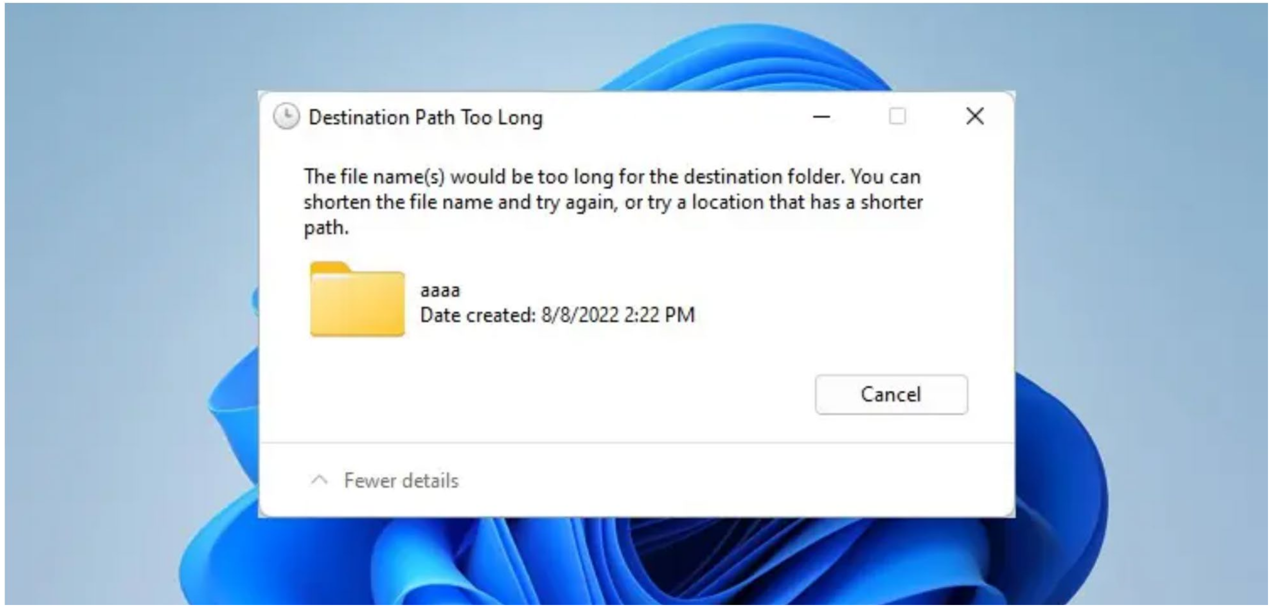
Data Loss: Accidental Loss

- Wiping out a disk drive containing information for an account worth $38 billion
- While doing routine maintenance work, the technician accidentally deleted applicant information for an oil-funded account—one of Alaska residents' biggest perks—and mistakenly reformatted the backup drive, as well.
- The department discovered its third line of defense, backup tapes, were unreadable.
- And the only backup was the paperwork itself—stored in more than 300 cardboard boxes.
- Cost: $220,700



Data Loss: Length of File Name

- Microsoft Windows has a MAX_PATH limit of ~256 characters.
  - If the length of the path and filename combined exceed ~256 characters you may not be able to delete/move/rename these paths/files.
- OneDrive, OneDrive for work or school and SharePoint: Limit is 400 characters in Microsoft 365.
  - Example: https://www.contoso.com/**sites/marketing/documents/Shared%20Documents/Promotion/Some%20File.xlsx**



Destination Path Too Long

The file name(s) would be too long for the destination folder. You can shorten the file name and try again, or try a location that has a shorter path.

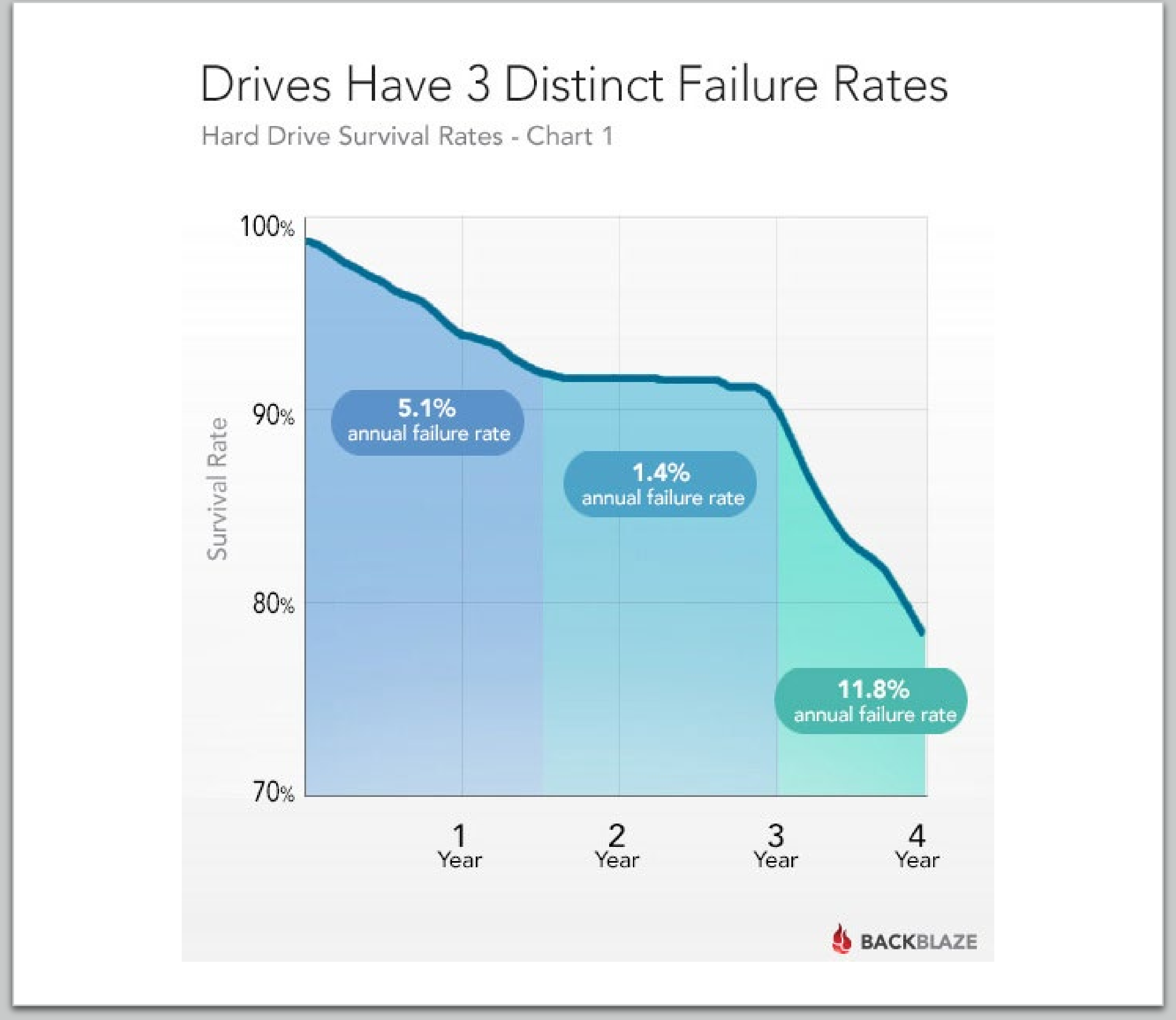
aaaa
Date created: 8/8/2022 2:22 PM

What percentage of hard drives fail within the first four years?

Less than 5%	
5%	
10%	
20%	
50%	
More than 50%	

Preventing Data Corruption

- Do backups
- MS Azure, OneDrive
- Google Drive (cloud storage)
- Reliable hardware
  - Backblaze kept up to 25,000 hard drives constantly online for four years. Every time a drive fails, they note it down, then slot in a replacement (2013 data)
  - 80% of drives last four years
  - https://www.backblaze.com/b2/hard-drive-test-data.html

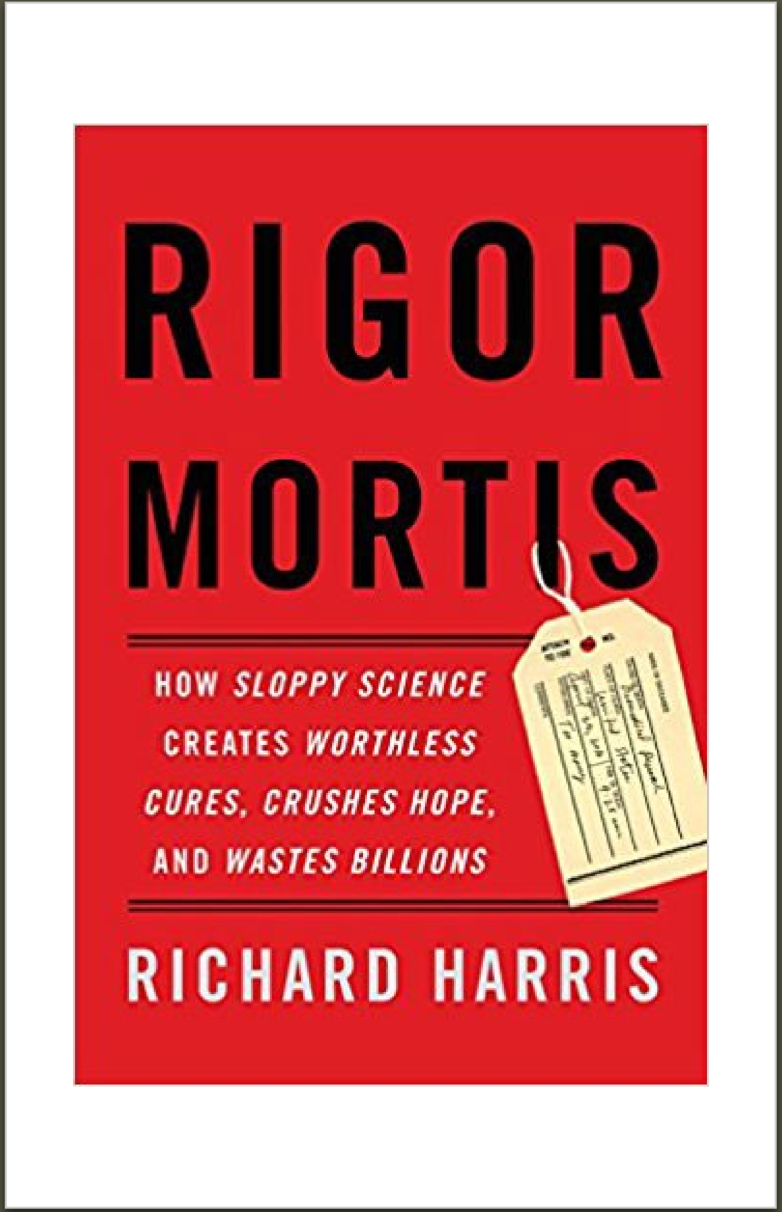


Section 3

Reproducibility, replicability, and research misconduct (Claudia)

Sloppy Science

- “Each year about a million biomedical studies are published in the scientific literature. And many of them are simply wrong.” (P. 7)
- “Scientists concerned about reproducibility broadly agree that fraud is not a major factor, but it does sit at the end of a spectrum of problems confronting biomedicine.” (P. 179)
- “Simply increasing transparency could go a long way toward reducing the reproducibility problems that plague biomedical research. […] The main project at the center is a data repository called the Open Science Framework.” (Pp. 145-146)
- Harris, Richard. Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions). Basic Books. Kindle Edition.


RIGOR MORTIS
HOW SLOPPY SCIENCE CREATES WORTHLESS CURES, CRUSHES HOPE, AND WASTES BILLIONS
RICHARD HARRIS

Reproducibility and Replicability

- Reproducibility
  - “the ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline”
- Replicability
  - “the chance that an independent experiment targeting the same scientific question will produce a consistent result”
- Source:
  - Leek, Jeffrey T., and Roger D. Peng. "Opinion: Reproducible research can still be wrong: Adopting a prevention approach." *Proceedings of the National Academy of Sciences* 112.6 (2015): 1645-1646.

Data Integrity

- Assurance of the accuracy and integrity of the data over its life cycle
  - No alterations when data are sent or received
  - No alterations between record updates
  - Storage on device that is highly reliable
- Example: Store genomics data as “read only” and back up on a device that is highly reliable

Achieving Reproducibility

- “Re-compute data analytic results given an observed dataset and knowledge of the data analysis pipeline”
  - Keep the data
  - Keep the analysis pipeline
    - Software versions
    - Parameter choices
- Teach your students how to manage the data for reproducibility
  - The final chapter of a thesis should have all the details on how someone else can use the data
  - Also important to respond to research misconduct allegations

The Toxic Lab, Replicability and Reproducibility, and Research Misconduct I

Charles Wood* wrote about two models of lab management, which "can both discourage trainees and encourage misconduct”:

> In the **executive model** of lab management, the principal investigator demands that trainees meet his or her expectations, often with a specific goal in mind. In its most toxic form, that goal can include specific experimental outcomes — so a trainee is told to do this experiment and get this particular result. [...] In the second toxic style of mentorship, the **competition model**, principal investigators give two or more trainees the same goal. The implication is that the one who completes the task first — or, more dangerously, the one who generates the data that conform best to the preconceived outcome — is the winner.

*Wood, C. Column: When lab leaders take too much control. Nature 491, 785–786 (2012). https://doi.org/10.1038/nj7426-785a (emphasis mine)

The Toxic Lab, Replicability and Reproducibility, and Research Misconduct II

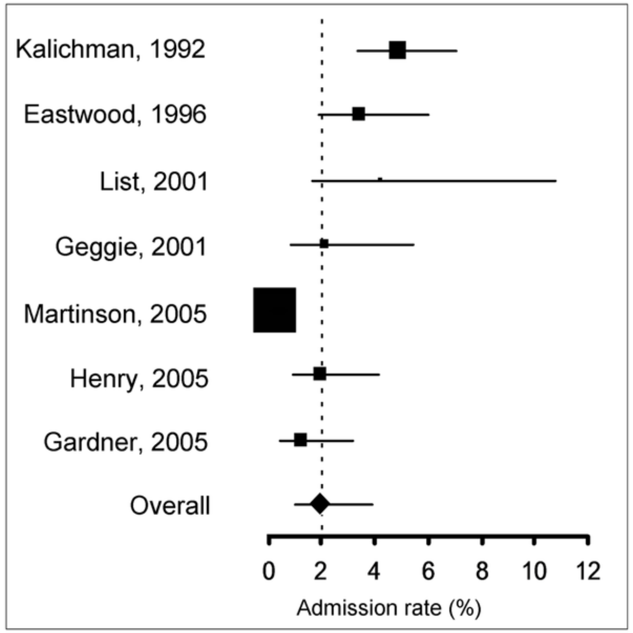
The climate and environment early career researchers encounter shape their ethical behavior, according to Mumford *et al*.*

> They identified five factors of environmental experiences: "[P]rofessional leadership, poor coping, lack of rewards, limited competitive pressure, and poor career direction," and four factors of climate perception: "[E]quity, interpersonal conflict, occupational engagement, and work commitment." They found that "it appears, at least among 1st-year doctoral students, that [environmental] experience exerts stronger effects on ethical decision making than the climate of the work group."

*Mumford, Michael D., et al. "Environmental influences on ethical decision making: Climate and environmental predictors of research integrity." Ethics & behavior 17.4 (2007): 337-366.

Research Misconduct: Prevalence

Figure 2. Forrest plot of admission rates of data fabrication, falsification and alteration in self reports.



- A pooled weighted average of 1.97% (N = 7, 95%CI: 0.86–4.45) of scientists admitted to have fabricated, falsified or modified data or results at least once –a serious form of misconduct by any standard– and up to 33.7% admitted other questionable research practices.
- **Data will be sequestered following accusation**

Fanelli D (2009) How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. PLOS ONE 4(5): e5738. https://doi.org/10.1371/journal.pone.0005738
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738

§ 93.105 Time Limitations

a) **Six-year limitation.** This part applies only to research misconduct occurring within six years of the date HHS or an institution receives an allegation of research misconduct.

b) Exceptions to the six-year limitation. Paragraph (a) of this section does not apply in the following instances:

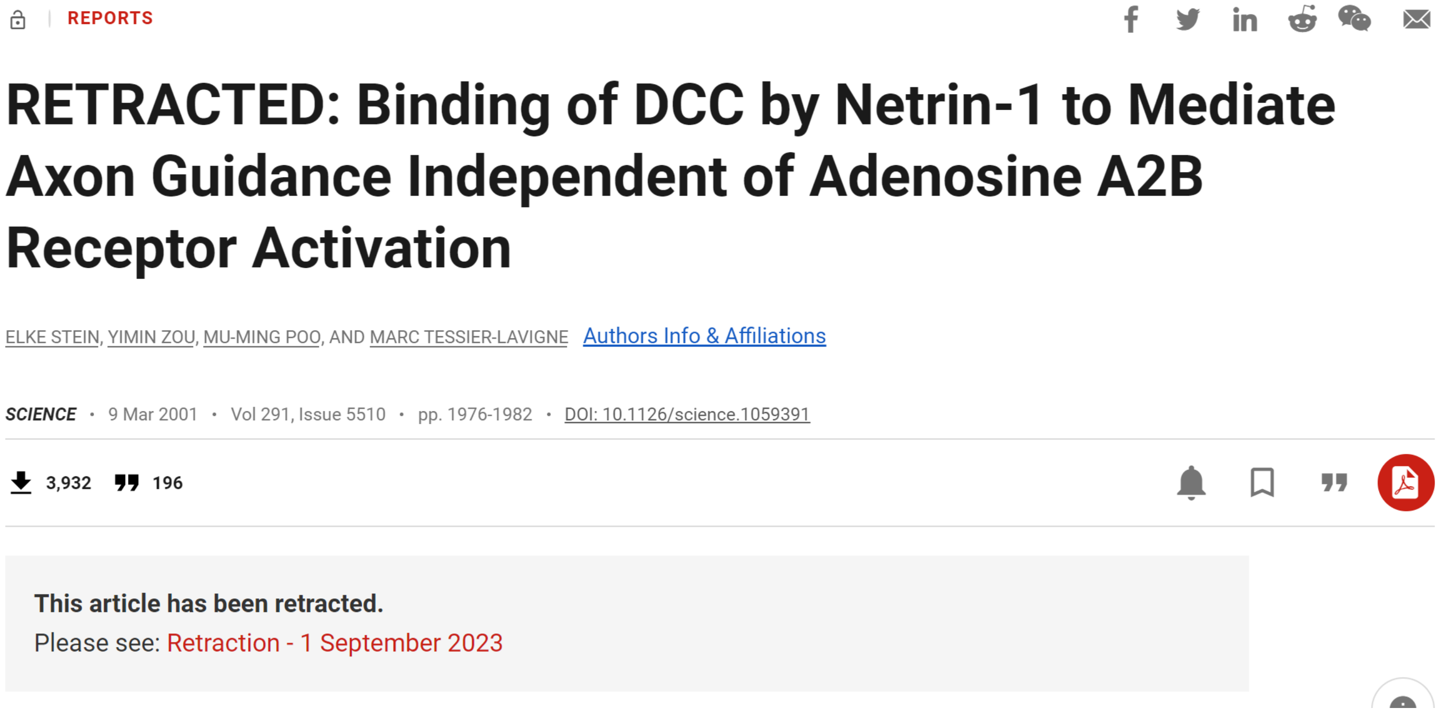
   1) **Subsequent use exception.** The respondent continues or renews any incident of alleged research misconduct that occurred before the six-year limitation through the **citation, republication or other use for the potential benefit of the respondent** of the research record that is alleged to have been fabricated, falsified, or plagiarized.

   2) **Health or safety of the public exception.** If ORI or the institution, following consultation with ORI, determines that the alleged misconduct, if it occurred, would possibly have a substantial adverse effect on the health or safety of the public.

c) “Grandfather” exception. If HHS or an institution received the allegation of research misconduct before the effective date of this part.

Research Misconduct: Manipulation of Western Blots

- Tessier-Lavigne, former president of Stanford, was involved in multiple allegations of image manipulations
- The Stanford investigation committee conclude that he did not have knowledge of the manipulation
- Tessier-Lavigne stepped down on August 31, 2023



REPORTS

RETRACTED: Binding of DCC by Netrin-1 to Mediate Axon Guidance Independent of Adenosine A2B Receptor Activation

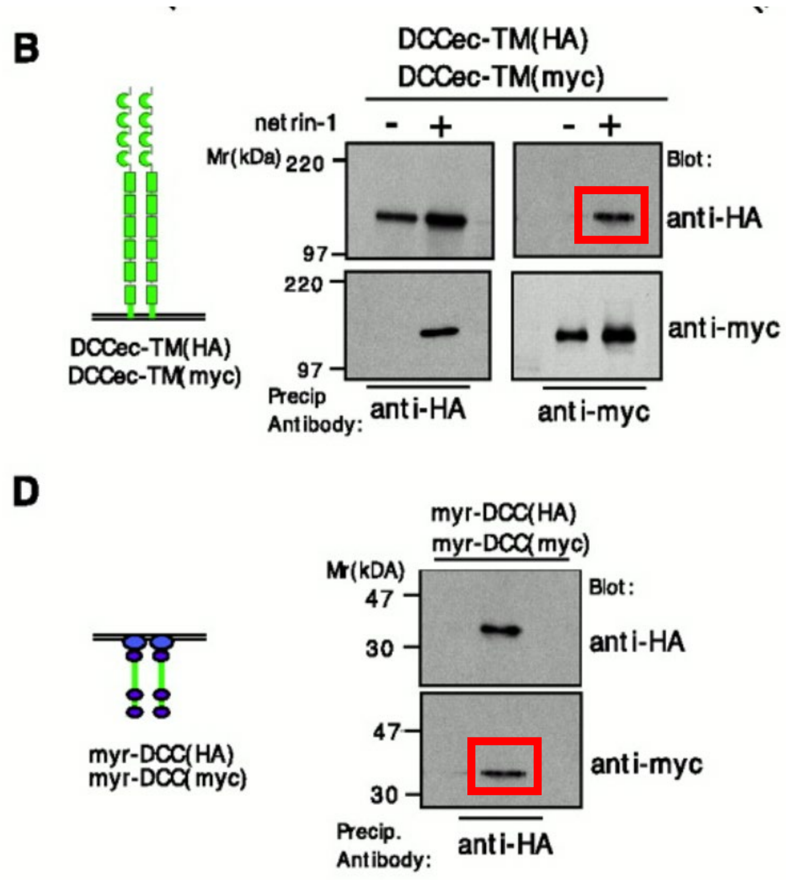
ELKE STEIN, YIMIN ZOU, MU-MING POO, AND MARC TESSIER-LAVIGNE Authors Info & Affiliations

SCIENCE • 9 Mar 2001 • Vol 291, Issue 5510 • pp. 1976-1982 • DOI: 10.1126/science.1059391

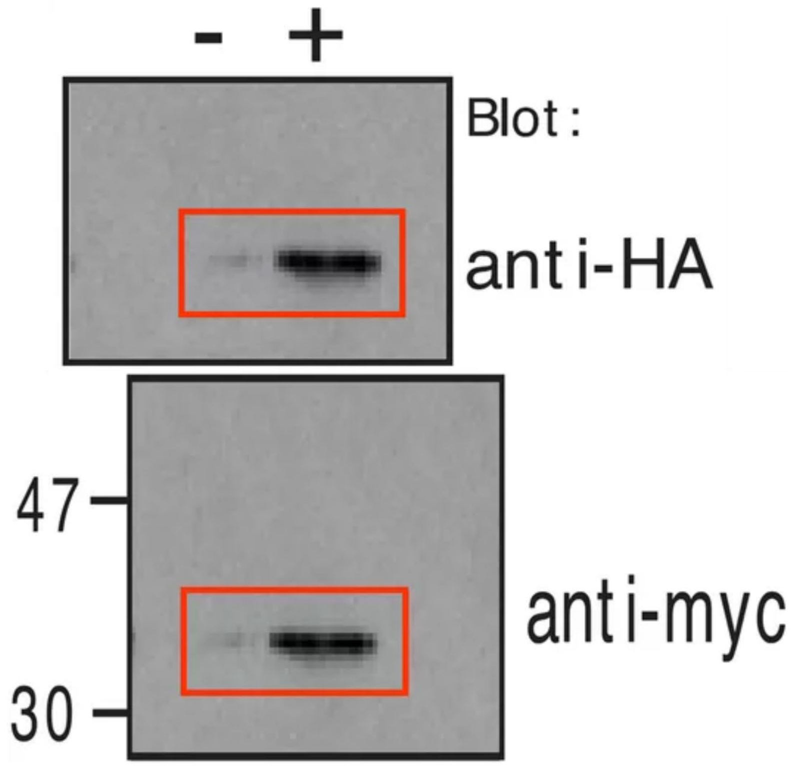
3,932 196

This article has been retracted.
Please see: Retraction - 1 September 2023

PubPeer: Figure 3B and D



After change in aspect ratio:



Source:
https://pubpeer.com/publications/59C3359E71EED451E01AF46CFDC0BC

ChatGPT, Hallucinations, and Plagiarism

- Retraction Watch: Withdrawn AI-written preprint on millipedes
  - Kahsay Tadesse Mawcha of Aksum University in Ethiopia, used ChatGPT to write the manuscript
    - Fake references that distort the effect of millipedes on crops
  - https://retractionwatch.com/2023/09/01/withdrawn-ai-written-preprint-on-millipedes-resurfaces-causing-alarm/

Wikipedia

Section 4

Basic Elements of a Data Management Plan (DMP) (Santi)

Have you written a data management plan?

YES	**NO**

Basic elements of a data management plan

Major components of all Plans

- What is generated?
- How is it securely handled?
- How is it maintained and accessed long-term?

+

- Roles and responsibilities
- Use of systems and platforms
- Documentation
- Security

Have you heard of the DMPTool?

YES	NO

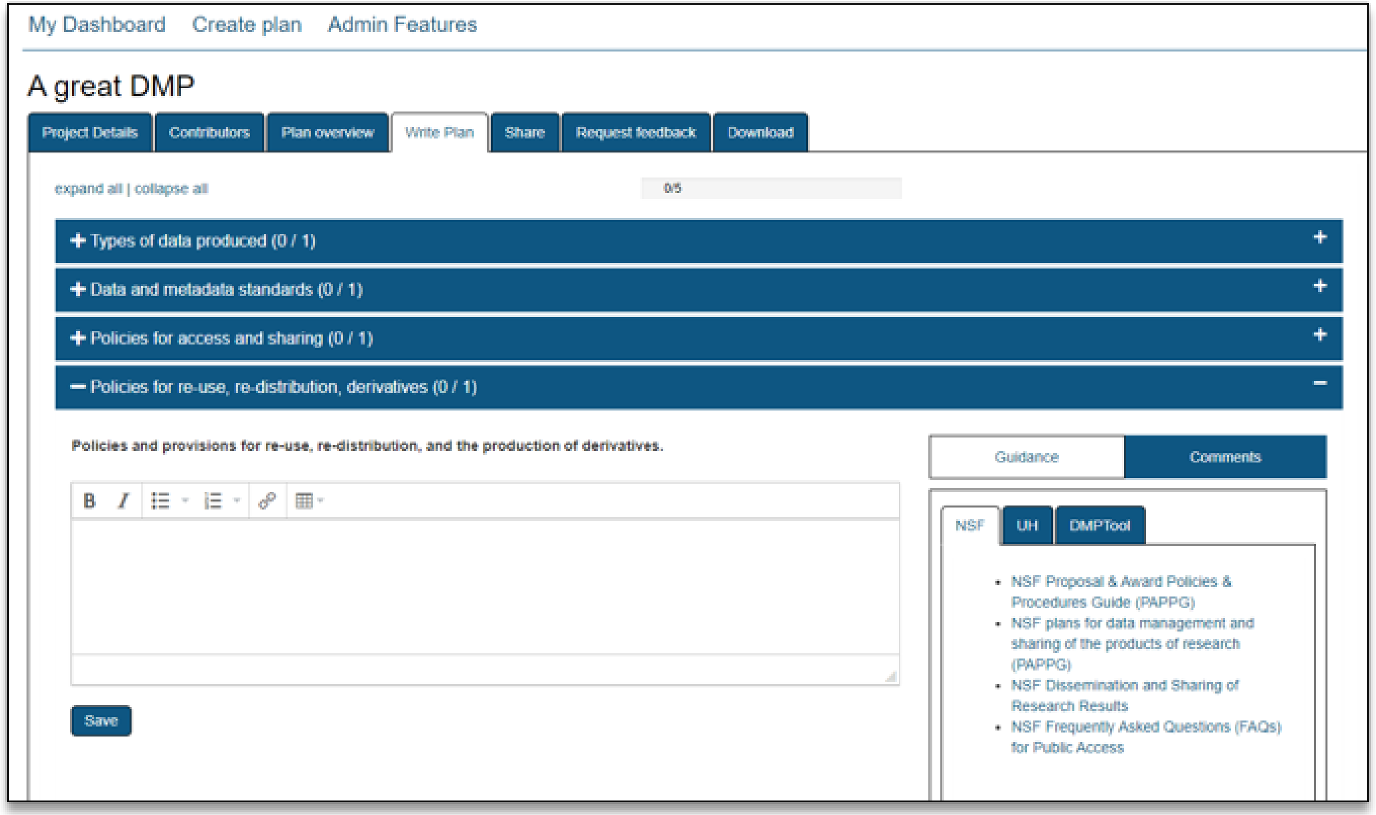
A recommended tool for writing

DMPTool

Build your Data Management Plan

https://dmptool.org

- Agency templates & guidance
- Institutional login
- Submit plans for feedback



A great DMP

+ Types of data produced (0 / 1)

+ Data and metadata standards (0 / 1)

+ Policies for access and sharing (0 / 1)

− Policies for re-use, re-distribution, derivatives (0 / 1)

Policies and provisions for re-use, re-distribution, and the production of derivatives.

- NSF Proposal & Award Policies & Procedures Guide (PAPPG)
- NSF plans for data management and sharing of the products of research (PAPPG)
- NSF Dissemination and Sharing of Research Results
- NSF Frequently Asked Questions (FAQs) for Public Access

The library as bookends for data management

Proposal

- Policy guidance
- DMP writing support
- Planning for post project needs
- Connections to resources

Post Project

- Archiving and Preservation
- Sharing
- Curation guidance
- UH Data Repository

Data Management Resources and contact information
https://guides.lib.uh.edu/datamanagement

Section 5

Data access, ownership, and security(Claudia)

Data Management and Sharing Policy

- MAPP 08.03.01
  - https://uh.edu/policies/_docs/mapp/08/080301.pdf
- SAM 07.A.08
  - https://uhsystem.edu/compliance-ethics/_docs/sam/07/7a8.pdf
- University of Houston: ownership (whether or not the research is externally funded)
- PI Responsibilities: stewardship, data management plan
- Colleges/Departments/Centers: resource providers
- Division of Research: policy owner, compliance
- Data storage, archiving, data sharing
- Level 1 data: Mission-critical information includes all research data obtained from third parties pursuant to an agreement or grant and/or other data necessary to substantiate research results or to satisfy grant-funding requirements, regardless of whether such data was developed by the university or obtained from third parties.

Who Can Access Your Data?

- Federal: Freedom of Information Act
  - *"The Freedom of Information Act (FOIA) is a law that gives you the right to access information from the federal government."*
  - Half of the FOIA requests for grants are by academic scientists who want to look at other researchers' grant applications
- Do Open Record Laws in Texas protect research data?
  - **Research data** produced by university faculty pursuant to a contract between the university and a third party is information that is collected, assembled, or maintained by a governmental body and that is connected to the transaction of official business. Consequently, the data is **public information** subject to the Open Records Act, Government Code chapter 552.
  - Section 51.914(1) of the Education Code deems **confidential** "scientific information . . . developed in whole or in part at a state institution of higher education" if the information has "a potential for being **sold, traded, or licensed for a fee**." Whether particular scientific information has a potential for being sold, traded, or licensed for a fee is a question requiring the resolution of fact issues. This office will therefore rely on the university's assertion that some of the requested information has this potential. Accordingly, the university must withhold certain of the requested information under section 51.914(1) of the Education Code as applied through section 552.101 of the Government Code.
    - Source: Office of the Attorney General of Texas, March 18, 1997: https://www2.texasattorneygeneral.gov/opinions/openrecords/48morales/ord/1997/htm/ord19970651.htm
  - Section 552.120 of the Public Information Act: Rare books, original manuscripts, personal papers, unpublished letters, and audio and video tapes held by an institution of higher education for the purposes of **historical research** are confidential, and the institution may restrict access by the public to those materials to protect the actual or potential value of the materials and the privacy of the donors.
    - Source: https://www.texasattorneygeneral.gov/sites/default/files/files/divisions/open-government/publicinfo_hb.pdf

Intellectual Property

- The Office of Technology Transfer and Innovation helps University employees with disclosures of inventions and patents/licensing
- Submit disclosure form to OTTI at least three months prior to public disclosure
- Public disclosure
  - Patent application needs to be filed within a year of a public disclosure
    - Posters
    - Presentations
    - Publications
    - All theses are publicly available
- Protect confidential information
  - Non-disclosure agreements
  - Marking confidential information as "Confidential"

What Is a Data Use Agreement (DUA)?

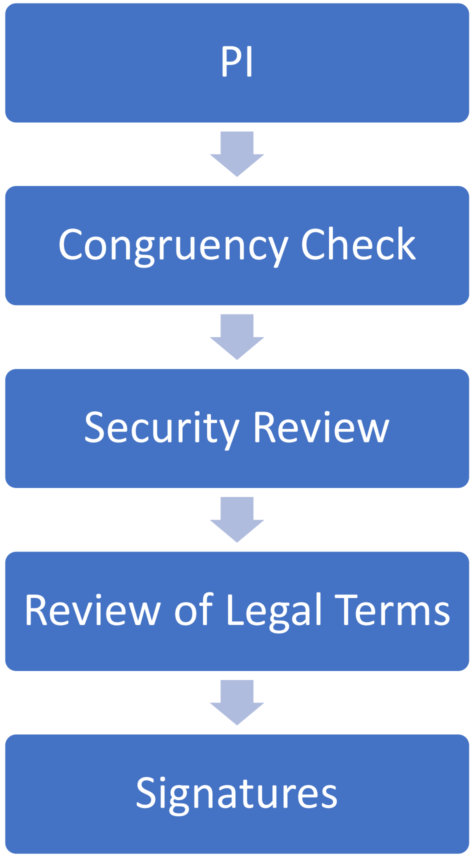
A **"limited data set"** is health information that is de-identified (it may include city; state; zip code; elements of date; and other numbers, characteristics, or codes not listed as direct identifiers).

A **Data Use Agreement** (DUA) is used for the transfer of data from nonprofit, government or private industry where the data are nonpublic or otherwise subject to restrictions on its use (e.g., limited data sets).

A data use agreement provides **satisfactory assurances** that the recipient of the data set will use or disclose the data set only for specified purposes and will only use approved hardware and software.

Components and Process

- Components
  - Data fields
  - Responsibilities of recipient
  - Permitted uses and disclosures
  - Term and termination
  - OCG link on how to go about getting a DUA: https://www.uh.edu/research/sponsored-projects/contracts/data-use-agreement/
- Breach
  - De-identified data are not subject to HIPAA
  - Unauthorized uses or disclosures of a limited data set may still constitute a 'breach'
  - The global average cost of a data breach in a 2019 study was $3.92 million (IBM Security/Ponemon Institute)



PI

Congruency Check

Security Review

Review of Legal Terms

Signatures

International Research and Travel

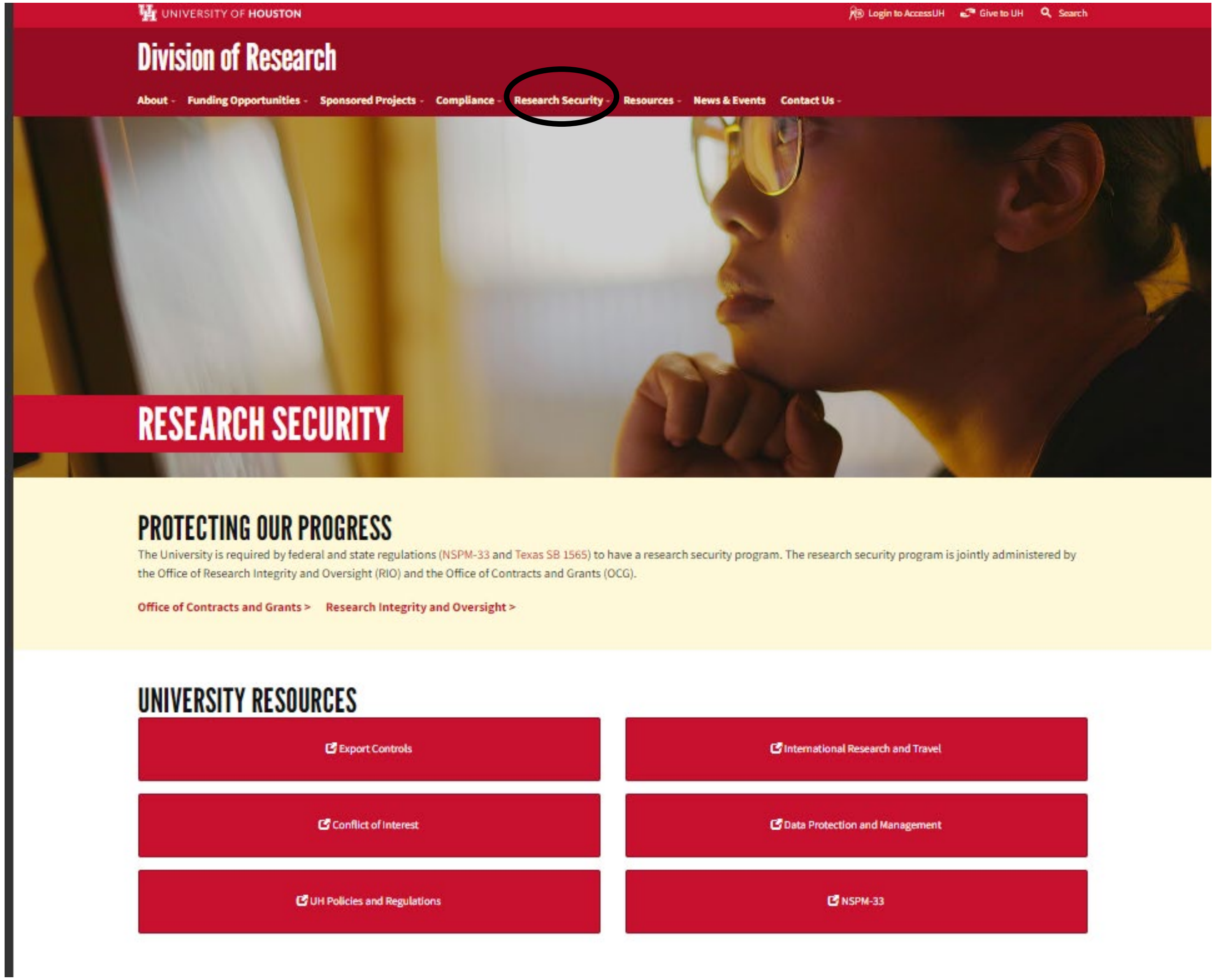
- Federal agencies require the disclosure of all sources of research support, foreign components, and financial conflicts of interest for senior/key personnel on research applications and awards. This includes support coming from foreign governments or other foreign entities.
- Foreign nationals are not allowed access to controlled technologies, data, and software (source code), even within the U.S. – “deemed export”
  - Sharing information about controlled technologies or sharing controlled technologies with foreign nationals falls under U.S. export control regulations.
  - This includes hosting foreign visitors to UH and visiting foreign countries.
- All employees who travel to destinations outside the United States must complete the Export Controls and Travel Embargo Form and receive approval from the Office of Contracts and Grants, if required, before leaving on the trip.
- Electronic devices must be “clean”
  - How clean depends on the status of the country /entity

Controlled Unclassified Information (CUI)

- CUI is unclassified information that requires safeguarding or dissemination controls, pursuant to and consistent with applicable law, regulations, and government-wide policies.
- Once a project is determined to be CUI it is managed under a Technology Control Plan and System Security Plan. This plan outlines the security measures researchers and staff must follow in order to protect the CUI data.
- Plan ahead **before** submitting proposals
  - The University cannot accept **all** awards
  - Include cost for hardware or cloud solution and managed IT services

A New Website under Construction

- Research Security
  - Export Controls
  - Conflict of Interest
  - UH Policies and Regulations
  - International Research and Travel
  - Data Protection and Management
  - NSPM-33



PROTECTING OUR PROGRESS

The University is required by federal and state regulations (NSPM-33 and Texas SB 1565) to have a research security program. The research security program is jointly administered by the Office of Research Integrity and Oversight (RIO) and the Office of Contracts and Grants (OCG).

Office of Contracts and Grants >   Research Integrity and Oversight >

UNIVERSITY RESOURCES

Export Controls
International Research and Travel
Conflict of Interest
Data Protection and Management
UH Policies and Regulations
NSPM-33

Section 6

Sharing your data: Open Access (Claudia)

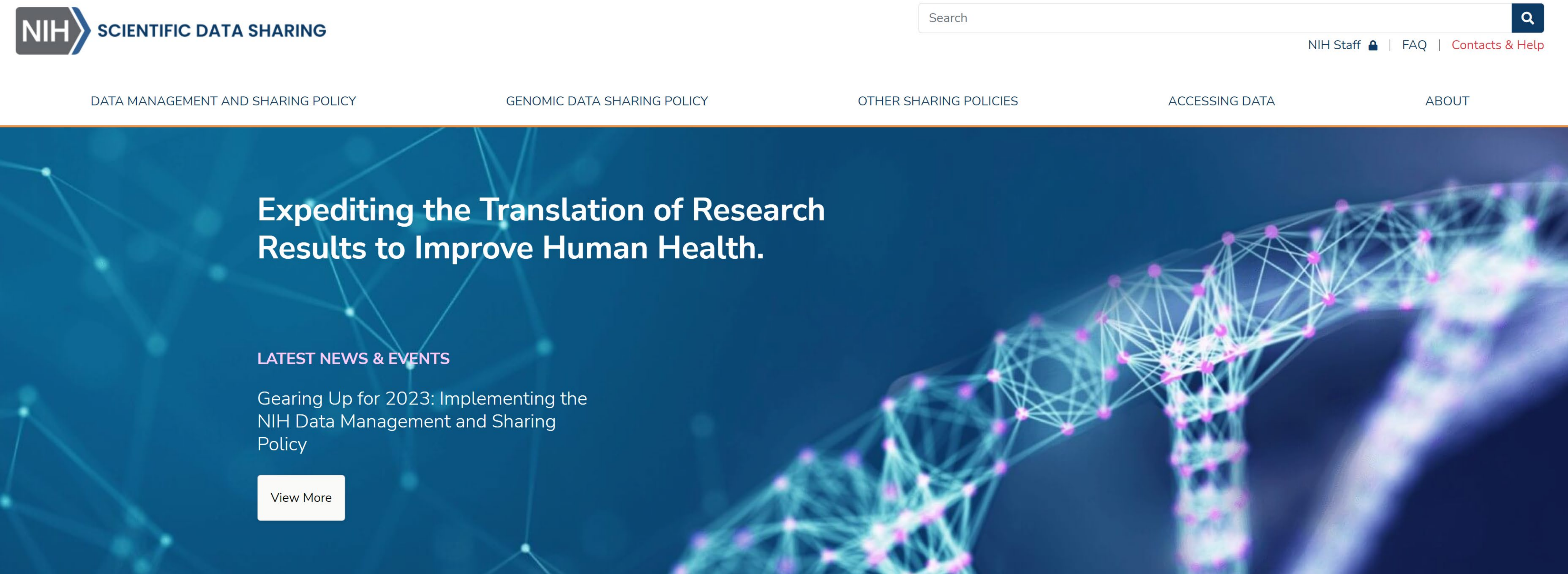
NIH Policy for Data Management and Sharing I

- Agencies will continue to put pressure on researchers to manage and share their data.
- Final NIH Policy for Data Management and Sharing: NOT-OD-21-013
    - Effective January 25, 2023
    - **Submission of a Data Management and Sharing Plan for all NIH-funded research** outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.
    - **Compliance with approved plan**
    - **Prior approval required for revisions to an approved DMS plan** (NOT-OD-23-185)
- Expectations
    - Data sharing should be maximized
    - Justifiable limits for technical/ethical/legal factors
    - Outline protection of privacy, rights, and confidentiality
    - Abide by existing laws, regulations, and policies
    - Plan prospectively for data management/sharing at all stages of the research process

NIH Policy for Data Management and Sharing II

- Submission and Review
  - Plan must be submitted with the application
  - Plan will be reviewed during assessment and NIH program staff will assess plans
  - Plan compliance through incorporation into Terms and Conditions and regular monitoring
  - Compliance may factor into future funding decisions
- Paying for Data Preservation on NIH Award
  - Up front, before award is over; all costs must be included in budget at application submission
- Electronic Research Notebook
  - Make sure that data standards in notebook are compatible with repository requirements
- Making Data Publicly Available
  - Findability required by plan
  - Persistent identifiers for data
  - Greater use of Data Use Agreements

NIH Policy for Data Management and Sharing III

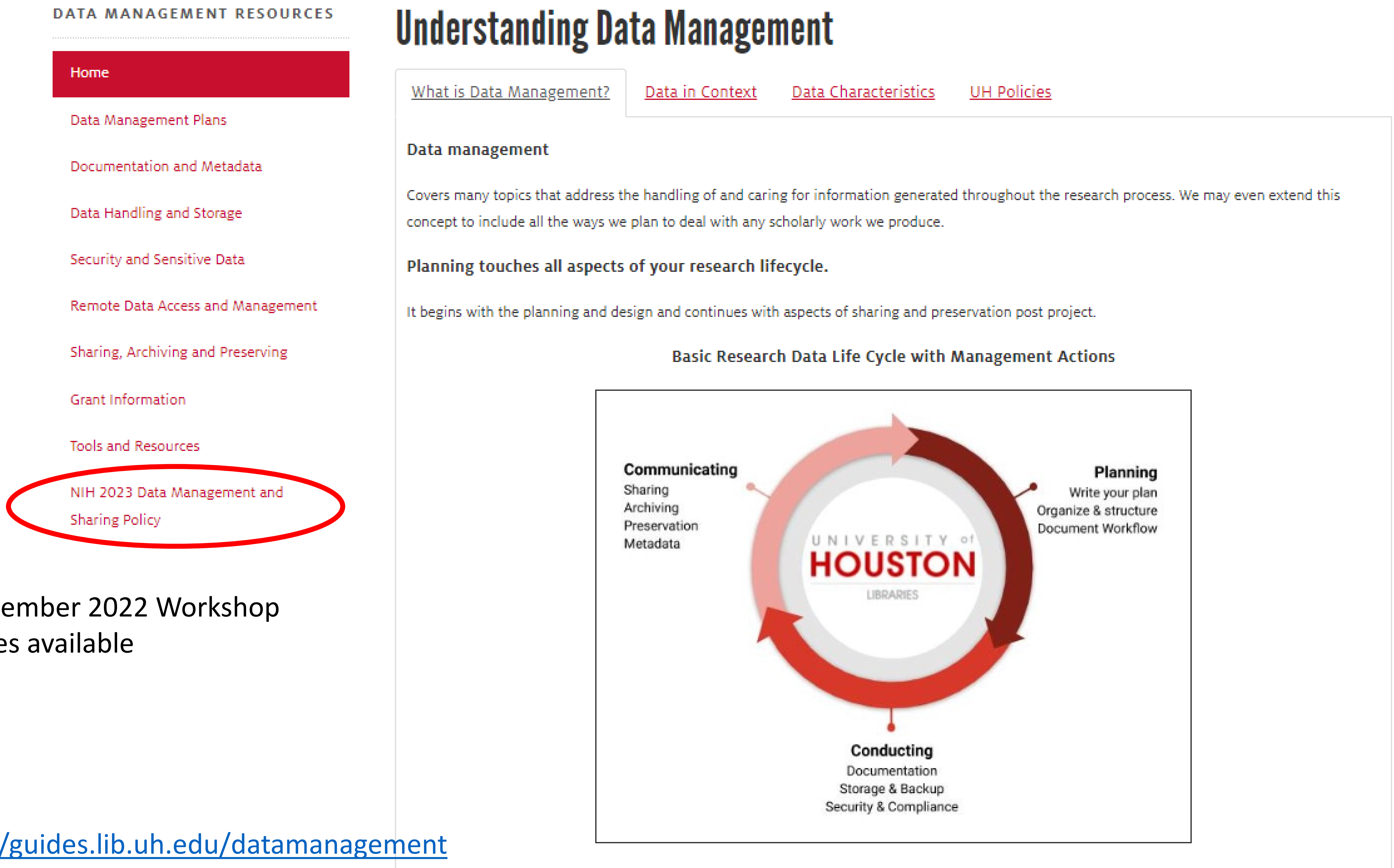

NIH SCIENTIFIC DATA SHARING
Search
NIH Staff
FAQ
Contacts & Help
DATA MANAGEMENT AND SHARING POLICY
GENOMIC DATA SHARING POLICY
OTHER SHARING POLICIES
ACCESSING DATA
ABOUT
View More

Expediting the Translation of Research Results to Improve Human Health.

LATEST NEWS & EVENTS

Gearing Up for 2023: Implementing the NIH Data Management and Sharing Policy

Website: https://sharing.nih.gov/
Video: https://www.youtube.com/watch?v=d2AdQZRjOHA&t=208s



DATA MANAGEMENT RESOURCES

- Home
- Data Management Plans
- Documentation and Metadata
- Data Handling and Storage
- Security and Sensitive Data
- Remote Data Access and Management
- Sharing, Archiving and Preserving
- Grant Information
- Tools and Resources
- NIH 2023 Data Management and Sharing Policy

Understanding Data Management

What is Data Management? | Data in Context | Data Characteristics | UH Policies

Data management

Covers many topics that address the handling of and caring for information generated throughout the research process. We may even extend this concept to include all the ways we plan to deal with any scholarly work we produce.

Planning touches all aspects of your research lifecycle.

It begins with the planning and design and continues with aspects of sharing and preservation post project.

Basic Research Data Life Cycle with Management Actions

November 2022 Workshop slides available

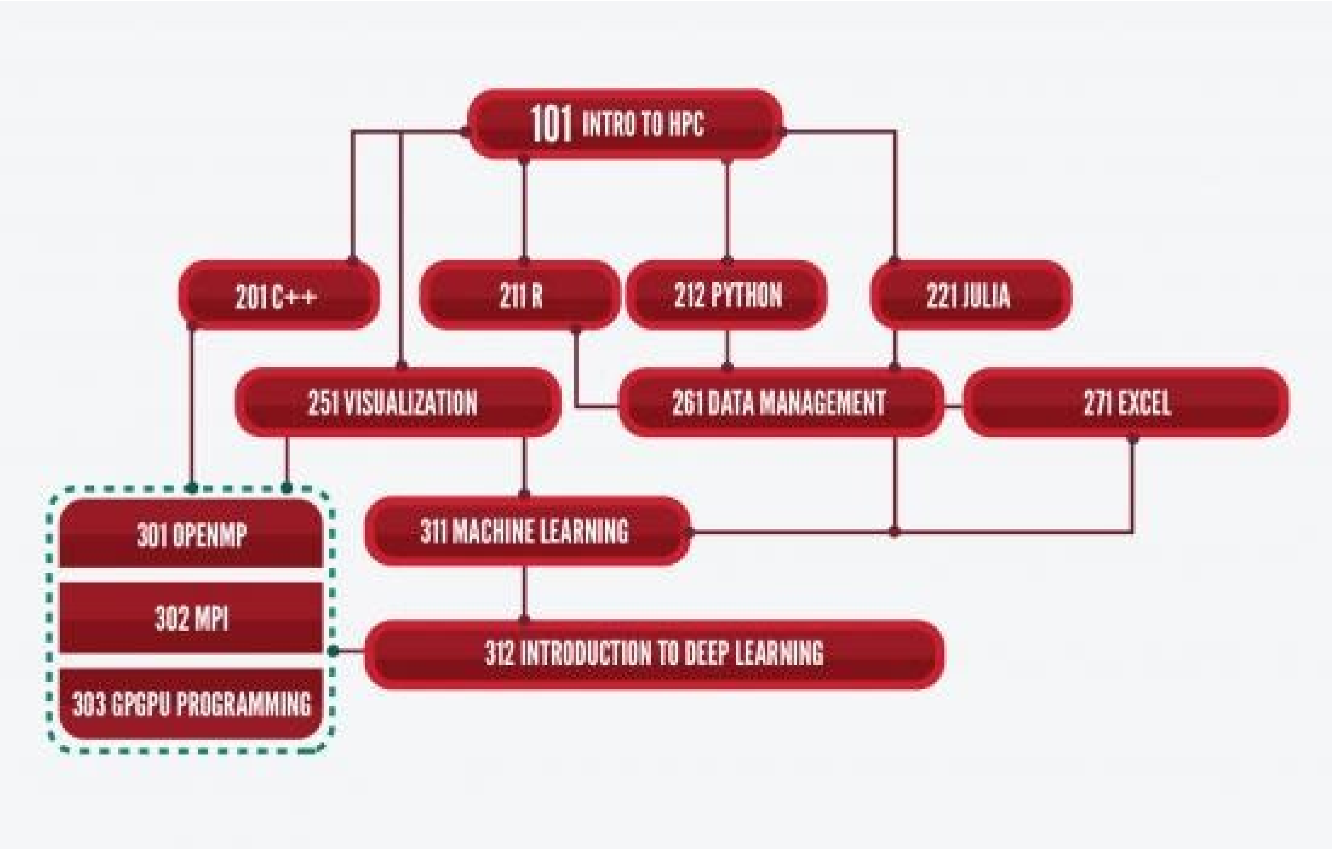
https://guides.lib.uh.edu/datamanagement

Open Access: Nelson Memo

- Learning from rapid dissemination of research during Covid: August 2022 Office of Science and Technology Policy released the “Nelson Memo”
  - Make federally funded research (peer reviewed publications and supporting scientific data) freely and publicly available without delay after publication
    - Publish as open access
    - Share final manuscript in preprint repositories
    - Use open data repositories
  - Applies to all federal agencies
- Steps
  - December 31, 2024: Deadline for agencies to publish final access policies
  - December 31, 2025: Deadline for new polices to become effective

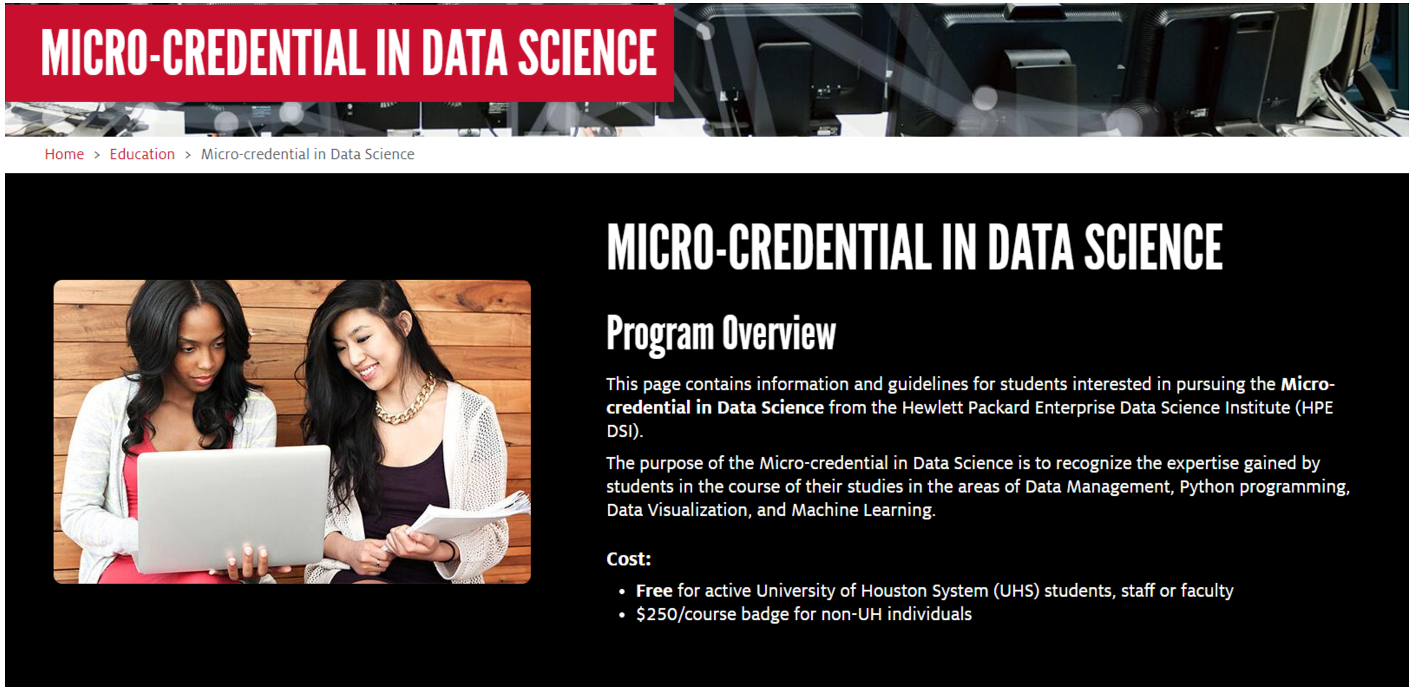
Hewlett Packard Enterprise
Data Science Institute
University of Houston

HPE DSI Courses: Learn how to manage and analyze data



Source: https://hpedsi.uh.edu/education/training

HPE DSI Data Science Microcredential



MICRO-CREDENTIAL IN DATA SCIENCE

Home > Education > Micro-credential in Data Science

MICRO-CREDENTIAL IN DATA SCIENCE

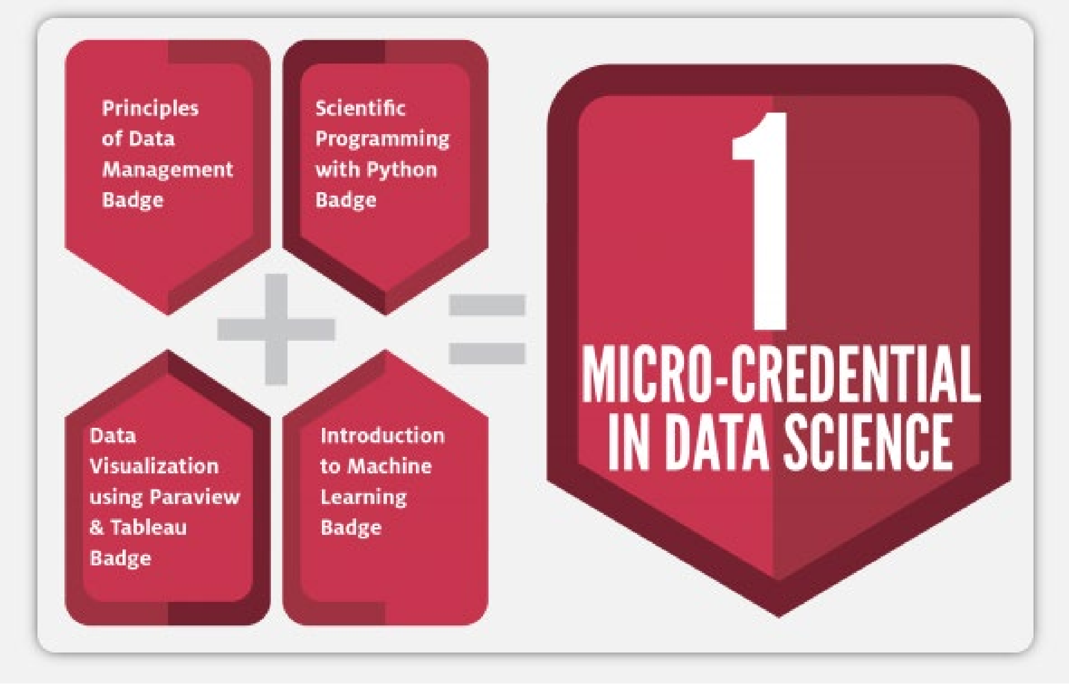
Program Overview

This page contains information and guidelines for students interested in pursuing the **Micro-credential in Data Science** from the Hewlett Packard Enterprise Data Science Institute (HPE DSI).

The purpose of the Micro-credential in Data Science is to recognize the expertise gained by students in the course of their studies in the areas of Data Management, Python programming, Data Visualization, and Machine Learning.

Cost:

- **Free** for active University of Houston System (UHS) students, staff or faculty
- $250/course badge for non-UH individuals



Principles of Data Management Badge

Scientific Programming with Python Badge

Data Visualization using Paraview & Tableau Badge

Introduction to Machine Learning Badge

1 MICRO-CREDENTIAL IN DATA SCIENCE

https://hpedsi.uh.edu/education/micro-credential-in-data-science

Section 7

Post Project (Santi)

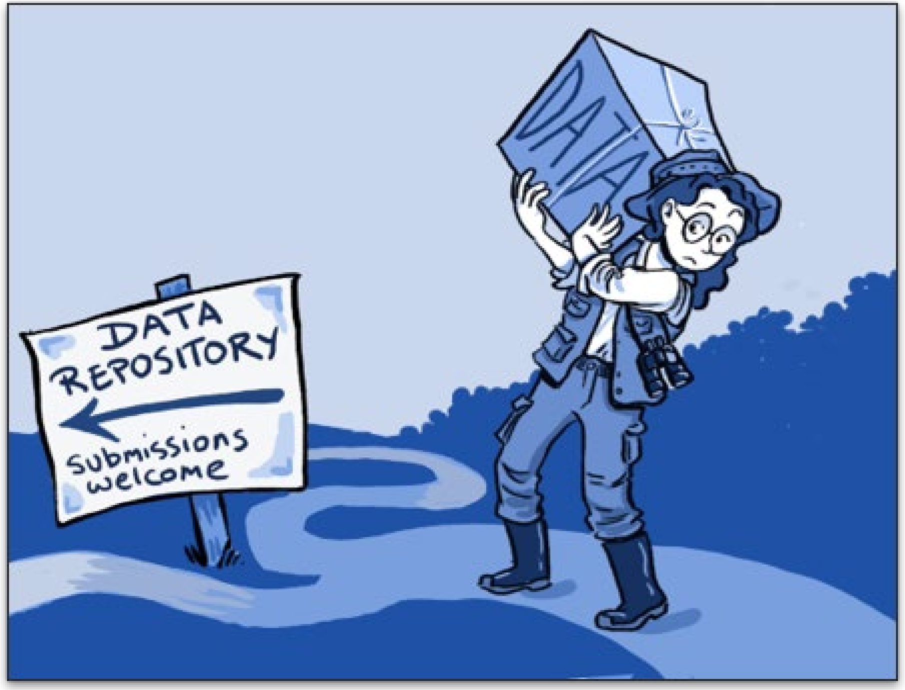
Poll 3: What are you currently doing with your data long-term?

I have not thought about this	
I don't have data	
My graduate students are responsible for this	
I have worked with the Libraries on this	
The data sit on my computer or external hard drive	
Other	

Archiving & Preservation vs. Storage

Benefits of a data repository

- File integrity and documentation
- Digital Object Identifiers and Cite-ability
- Greater discovery and access
- Less pain and time for you
- Federal grant policies



UH library offers post-project support for data and related materials.

Plan at the outset for long term access: Budget, Data format choice, Documentation

Choosing a repository

- What is required by funders and publishers?
- Where is my research community depositing?
- Do the parameters of the repository fit my data?

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

UH Data Repository

Archive, Preserve, and Share

- Open Access
- Free to all UH researchers
- Digital Object Identifier & Citation
- Up to 10 GB per project
- Local support

COUGAR
ROAR

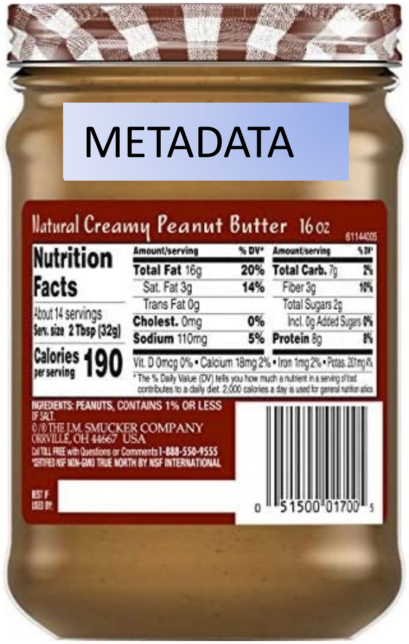
Texas Data Repository
University of Houston

https://dataverse.tdl.org/dataverse/uh

Have you heard about metadata?

YES	NO

What is Metadata?


METADATA
Natural Creamy Peanut Butter 16 oz
61144005
Nutrition Facts
About 14 servings
Serv. size 2 Tbsp (32g)
Calories per serving 190
Amount/serving
% DV*
Total Fat 16g 20%
Sat. Fat 3g 14%
Trans Fat 0g
Cholest. 0mg 0%
Sodium 110mg 5%
Total Carb. 7g 2%
Fiber 3g 10%
Total Sugars 2g
Incl. 0g Added Sugars 0%
Protein 8g 0%
Vit. D 0mcg 0% • Calcium 18mg 2% • Iron 1mg 2% • Potas. 201mg 4%
* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.
INGREDIENTS: PEANUTS, CONTAINS 1% OR LESS OF SALT.
©/® THE J.M. SMUCKER COMPANY
ORRVILLE, OH 44667 USA
Call TOLL FREE with Questions or Comments 1-888-550-9555
*CERTIFIED NSF NON-GMO TRUE NORTH BY NSF INTERNATIONAL
BEST IF USED BY:
0 51500 01700 5

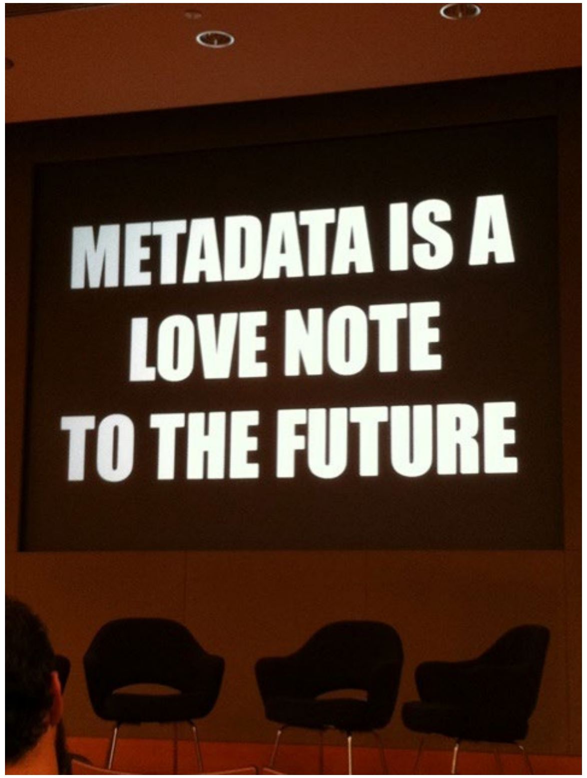
Descriptive information about the data, its creation, and use.
Variables, units of measure, standards, codes, procedures...

Data Curation & Documentation for Reproducibility

Choosing what to make accessible...

Needed:

- File management & conventions throughout a project
- Descriptive and structural metadata (Documentation)
    - What's in the files & the relationship between them
    - Apply standards where applicable and/ or use readme files



Funder focus on sharing and access requirements is connected to quality documentation and capturing related research materials beyond the data.

Questions?

Contact Information

Claudia Neuhauser
cmneuhauser@uh.edu

Santi Thompson
sathompson3@uh.edu