From Clinic to Analysis: Why Setting Up and Understanding Your Metadata Matters

Presented by:

Nicole Jimenez, PhD

Department of Obstetrics and Gynecology Postdoc

2022- 2023 Data Science Fellow

03-24-2023

Agenda

- Temperature Check
- Data Science/ Open Science
- Data Science Initiatives
- Metadata
- Study Design for Downstream Analysis
- Metadata Examples
- Public Databases
- Inspecting Data
- Tools, Tips, Resources -> I'll send slides so you have these!

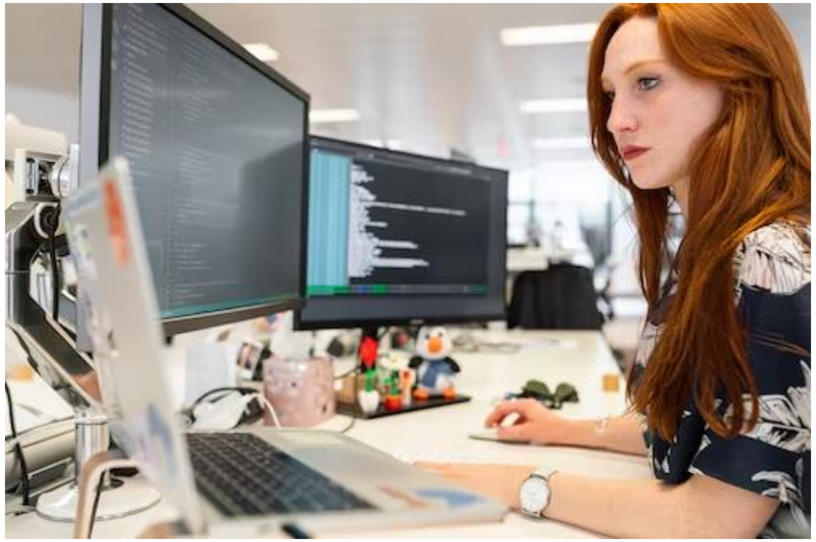
Gauge: Where does the audience background?

Basic Science Researcher (1)

Translational Data Scientist (3)

Clinician/Clinical Team Researcher (2)

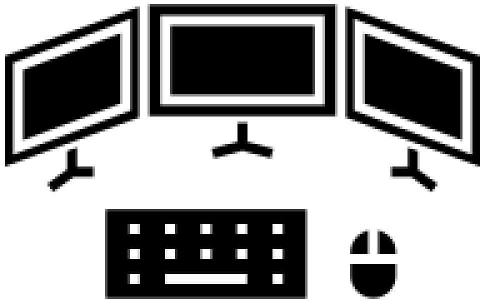
Software Developer/
Informatics Guru (4)



Gauge: Where does the audience stand on DS?

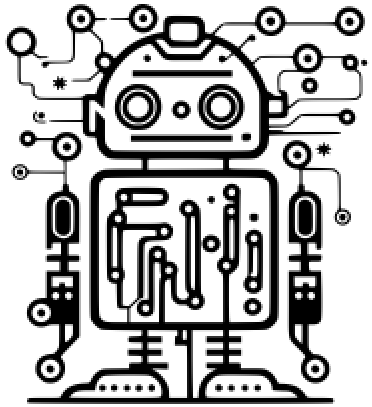
No background (1)

More than Average/ Still learning (3)



Some/ Minimal (2)

I dream in Code and Data (4)



Gauge: Do you have Datasets to analyze or will be analyzing

Learning/Planning Project (1)

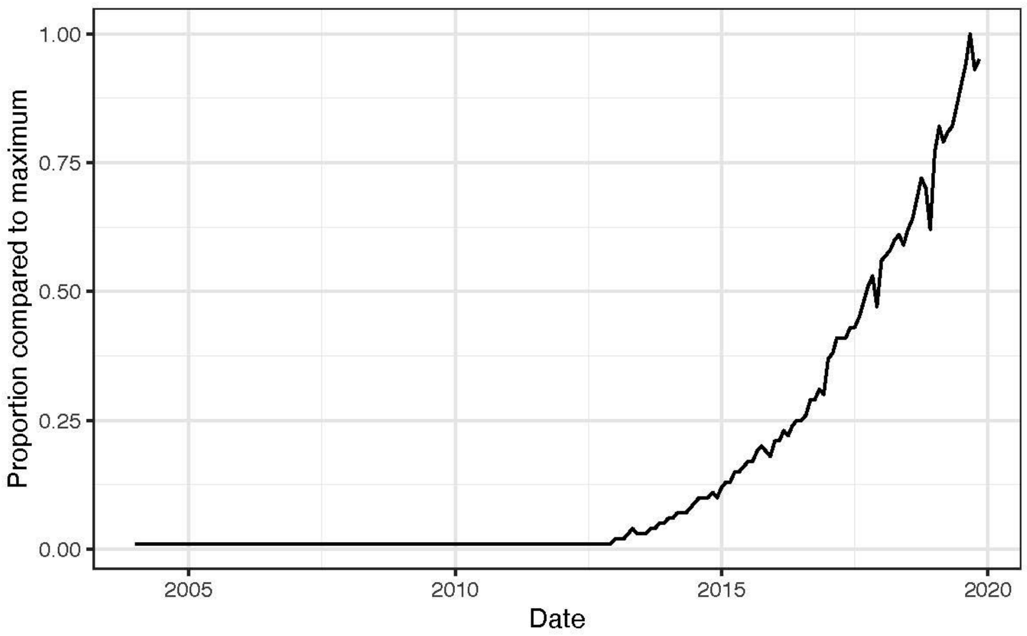
Recruitment/Data Collection(2)

Utilizing Public Data/ Data Organization (3)

I have Tons of Data and Need to Analyze ASAP (4)

What is Data Science?

Trends in Google searches of Data Science



‘data science is an umbrella term to describe the entire complex and multistep processes used to extract value from data.’ - Jeannette Wing , 2019

Backend data science*:

- Hardware
- Efficient computing
- Data storage infrastructure
- Data engineering

Frontend data science*:

- Data analysis
- Machine learning engineers
- Deep learning developers

*may depend on field

Biomedical Research Data:

•Fundamental research using model organisms (such as mice, fruit flies, and zebrafish)

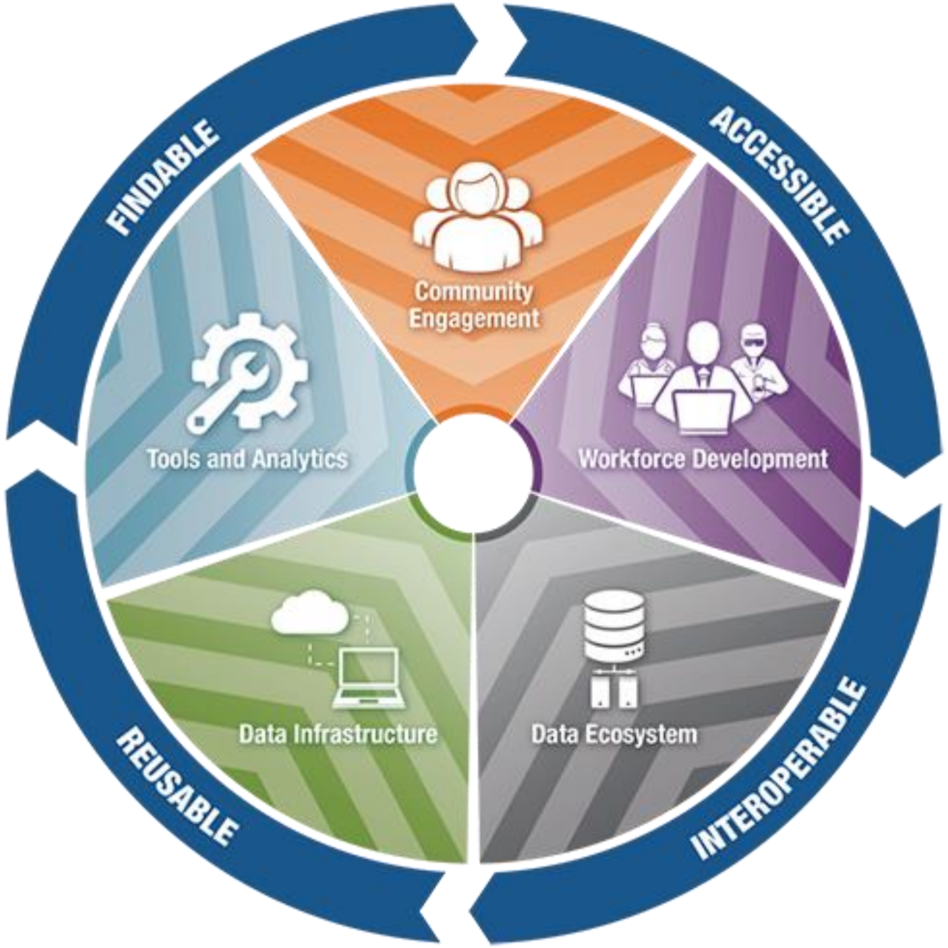
•Clinical studies (including medical images)

•Observational and epidemiological studies (including data from electronic health records and wearable devices).

Creation of a data-driven workforce; the world is data!

https://hdsr.mitpress.mit.edu/pub/gg6swfqh#n7oh8szsv6y

Open Science



Open Science (OS) is the movement to make scientific research, data and their dissemination available to any member of an inquiring society, from professionals to citizens. - **O**pen **R**esponsible research and **I**nnovation to further **O**utstanding k**N**owledge (**ORION**)

What is FAIR?

Biomedical research data should adhere to FAIR principles, meaning that it should be Findable, Accessible, Interoperable, and Reusable.

- To be **Findable**, data must have unique identifiers, effectively labeling it within searchable resources.
- To be **Accessible**, data must be easily retrievable via open systems and effective and secure authentication and authorization procedures.
- To be **Interoperable**, data should "use and speak the same language" via use of standardized vocabularies.
- To be **Reusable**, data must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable "owner's manual," or provenance.

https://www.orion-openscience.eu/
https://datascience.nih.gov/about/odss

https://howtofair.dk/

Data Science Initiatives in Biomedical Research

NIH National Institutes of Health
Office of Data Science Strategy

Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
• Optimize data storage and security • Connect NIH data systems	• Modernize data repository ecosystem • Support storage and sharing of individual datasets • Better integrate clinical and observational data into biomedical data science	• Support useful, generalizable, and accessible tools and workflows • Broaden utility of and access to specialized tools • Improve discovery and cataloging resources	• Enhance the NIH data-science workforce • Expand the national research workforce • Engage a broader community	• Develop policies for a FAIR data ecosystem • Enhance stewardship

Figure 2. NIH Strategic Plan for Data Science: Overview of Goals and Objectives

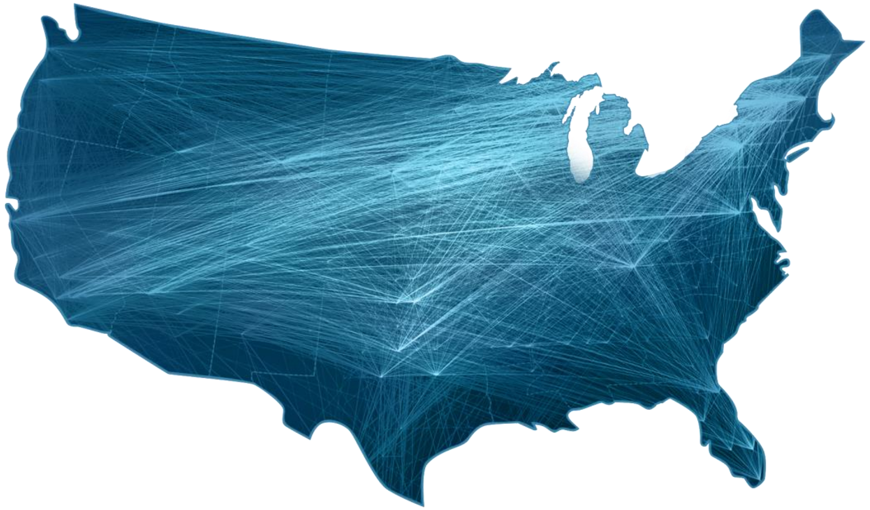
[NIH Data Science Strategic Plan]

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

Open Knowledge Roadmap

Pathways to Enable Open-Source Ecosystems

Experiential Learning for Emerging and Novel Technologies (ExLENT) program



https://beta.nsf.gov/tip/latest

Data Science Initiatives in Biomedical Research

College of Medicine Phoenix

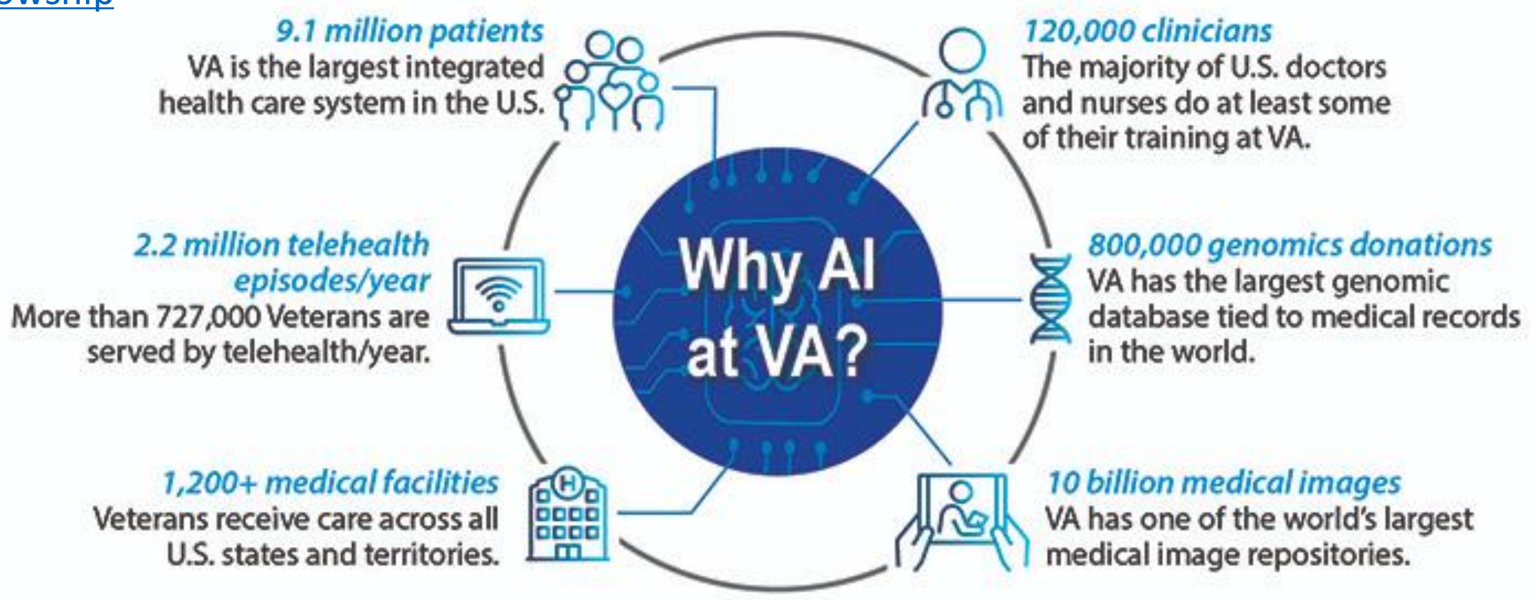
The Division of Clinical Analytics and Decision Support (CDADS) provides timely, targeted, meaningful data analytics and clinical decision support systems (CDSS) for the purposes of research, clinical care and quality improvement at the University of Arizona College of Medicine – Phoenix and associated institutions. https://phoenixmed.arizona.edu/clinicaldata

Biomedical Informatics - Clinical Fellowship Program is designed to effectively prepare fellows for a career in clinical informatics by providing them with the knowledge, skills and attitudes to support informatics-enabled improvement of clinical services.

https://phoenixmed.arizona.edu/clinical-informatics-fellowship

RESEARCH, INNOVATION & IMPACT
Data Science Institute

Data Science Institute aims to foster the next generation of data-driven research by encouraging university-wide interdisciplinary collaboration, gaining external visibility, developing industry alliances, and increasing funding for research at the University of Arizona (UA). https://datascience.arizona.edu/



https://www.research.va.gov/naii/default.cfm

https://www.research.va.gov/naii/BD-STEP/

What is Metadata?

Metadata, **“data about data,”** provides information such as data content, context, and structure, which is also valuable to the biomedical research community as it affects the ability of data to be found and used. - NIH

One definition commonly applied to the concept of metadata is the simple phrase **"data about data."** This simplistic definition, however, belies the significance and complexity of the nature of metadata. - American Health Information Management Association

What comes to mind when you hear the word metadata?

https://library.ahima.org/doc?oid=106378#.ZBy0PHbMK5d

[NIH Data Science Strategic Plan]()

Types of Metadata?

- **Administrative metadata:** data about a project or resource that are relevant for managing it; E.g. project/resource owner, principal investigator, project collaborators, funder, project period, etc. They are usually assigned to the data, before you collect or create them.

- **Descriptive or citation metadata:** data about a dataset or resource that allow people to discover and identify it; E.g. authors, title, abstract, keywords, persistent identifier, related publications, etc.

- **Structural metadata:** data about how a dataset or resource came about, but also how it is internally structured. E.g. the unit of analysis, collection method, sampling procedure, sample size, categories, variables, etc. Structural metadata have to be gathered by the researchers according to best practice in their research community and will be published together with the data.

https://carpentries-incubator.github.io/fair-bio-practice/05-intro-to-metadata/index.html

Types of Metadata? - Exercise

What are the types of metadata observed?

	Assembly	GenBank	Scientific name ↑	Size (…	Date	BioProject	Genes
☐	ASM1888363v1	GCA_018883635.1	Candidatus Lactobacillus …	1.748	Jul, 2021	PRJNA543206	1,856
☐	BRZ_IL__bin99	GCA_944327185.1	Candidatus Lactobacillus …	1.546	Jul, 2022	PRJEB53581	
☐	ASM1888367v1	GCA_018883675.1	Candidatus Paralactobaci…	1.161	Jul, 2021	PRJNA543206	1,178
☐	ASM883148v1	GCA_008831485.1	Lactobacillus acetotolera…	1.684	Nov, 2019	PRJNA566216	1,658

https://carpentries-incubator.github.io/fair-bio-practice/05-intro-to-metadata/index.html

Metadata Lifecycle

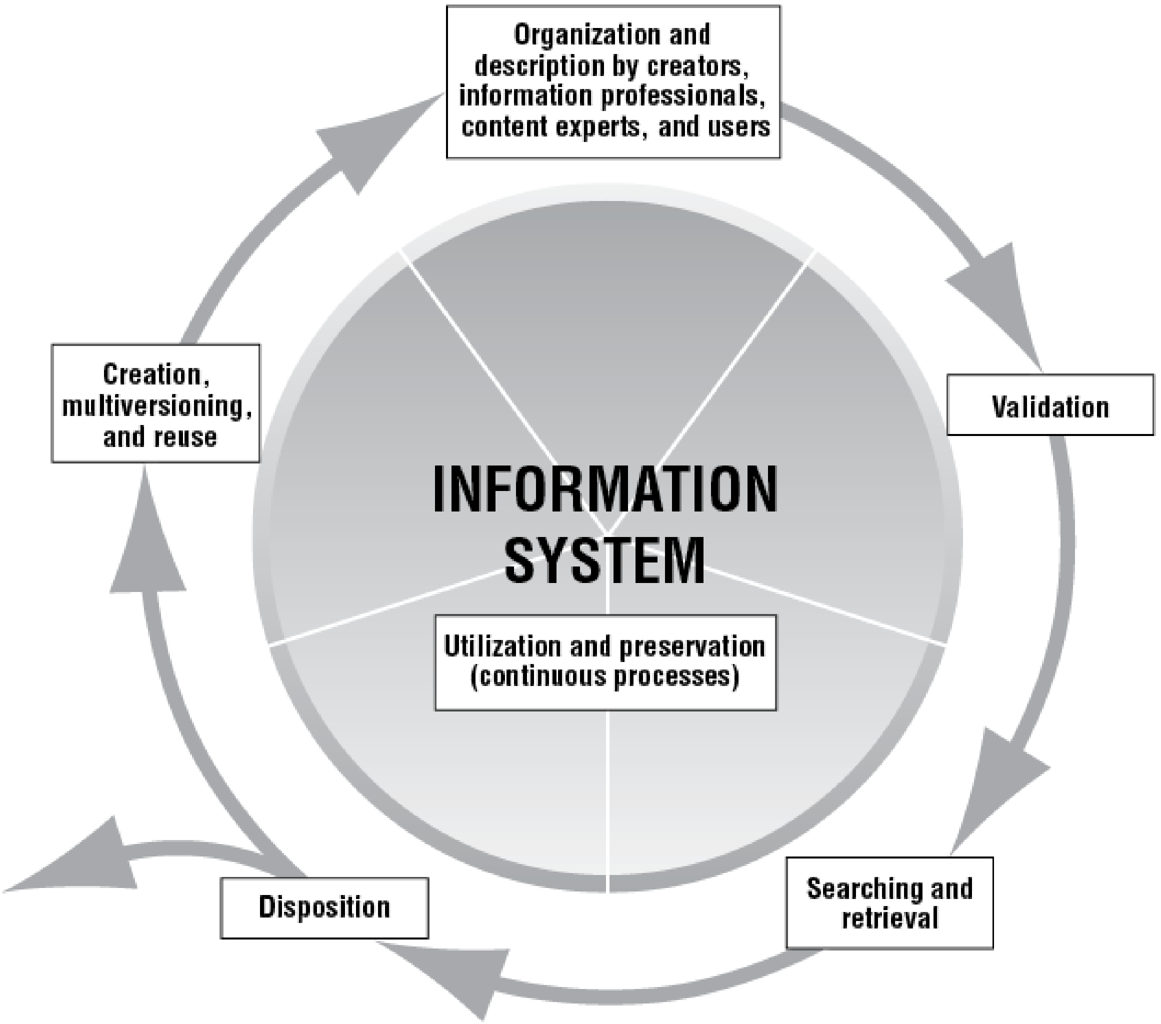


Table 2: Metadata in the Lifecycle

Information Lifecycle Phase	Example Metadata
Creation/Generation	Source, Date created, Time entered, Author, Version number
Classify/Index	File name, Document name
Store/Maintain	Last date accessed, Date archived
Search/View/Share	Patient name, Record number
Secure/Disclosed	Date disclosed, Party disclosed, Status of disclosure
Retain/Preserve/Dispose	Retention date, Disposition of data, Status of record (hold/active/inactive)

https://library.ahima.org/doc?oid=106378#.ZBy0PHbMK5d

Standardization

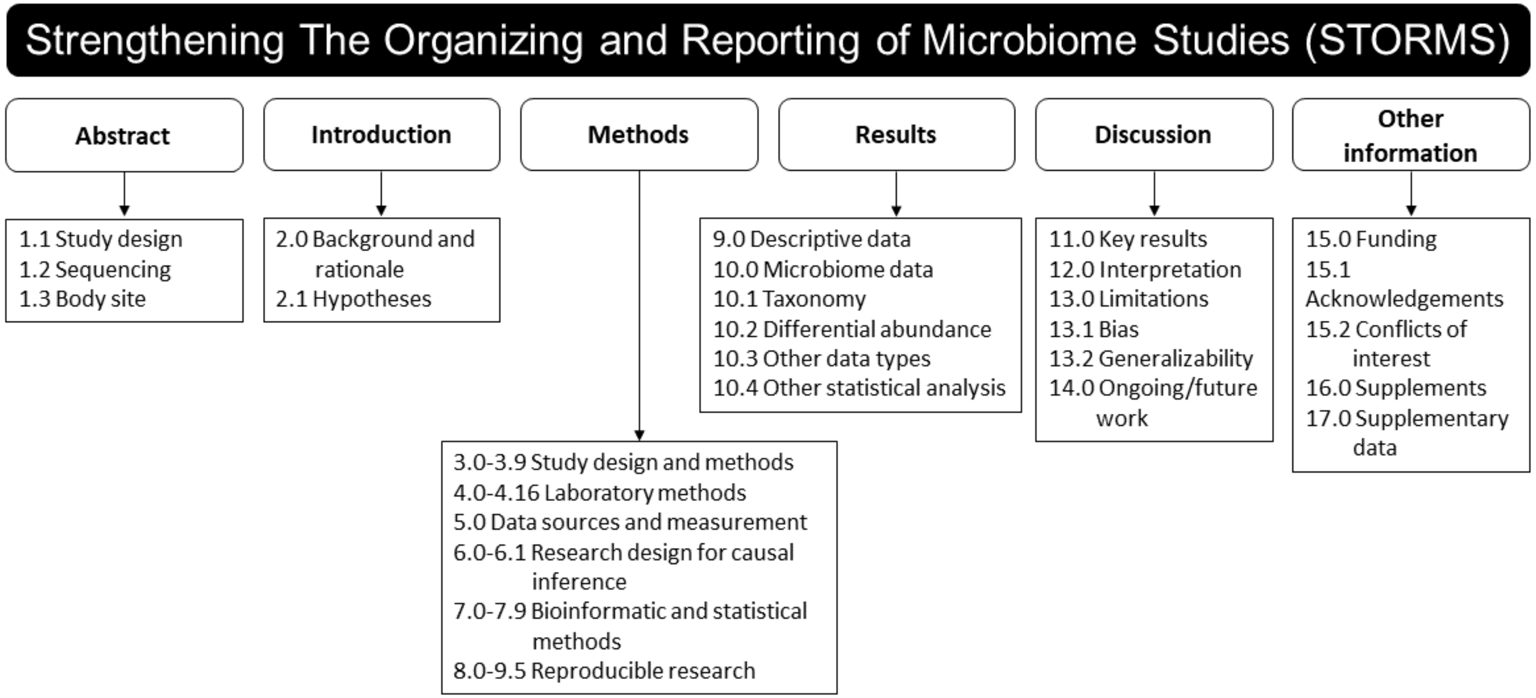
The **minimum information standard** is a set of guidelines for reporting data derived by relevant methods in biosciences. If followed, it ensures that the data can be easily verified, analyzed and clearly interpreted by the wider scientific community. Keeping with these recommendations also facilitates the foundation of structuralized databases, public repositories and development of data analysis tools. Individual minimum information standards are brought by the communities of cross-disciplinary specialists focused on issues of the specific method used in experimental biology.

Standardization of methods, data collection techniques or reporting will ultimately help with better metadata for easier cross-comparison studies or to identify datasets that best meet your research questions.

https://www.dcc.ac.uk/guidance/standards/metadata

https://carpentries-incubator.github.io/fair-bio-practice/05-intro-to-metadata/index.html

Standardization



What standards on data collection/metadata reporting does your field have?

https://www.stormsmicrobiome.org/

https://datamanagement.hms.harvard.edu/collect-analyze/documentation-metadata

Plan Ahead -> "Data Management"

- What is your study question?
- Will there be other factors you need to account for?
- How are you collecting your data?
- Will it be interpretable by computers? -> Interoperable to others
- How will you store your data?

All will help you organize metadata and data!

Try not throw the kitchen sink at your research cohort or data collection, this will ultimately lead to more time cleaning the data and take more thought on how to utilize, create stewardship for this data, and best report the data.

https://datacarpentry.org/wrangling-genomics/01-background/

Plan Ahead -> Example

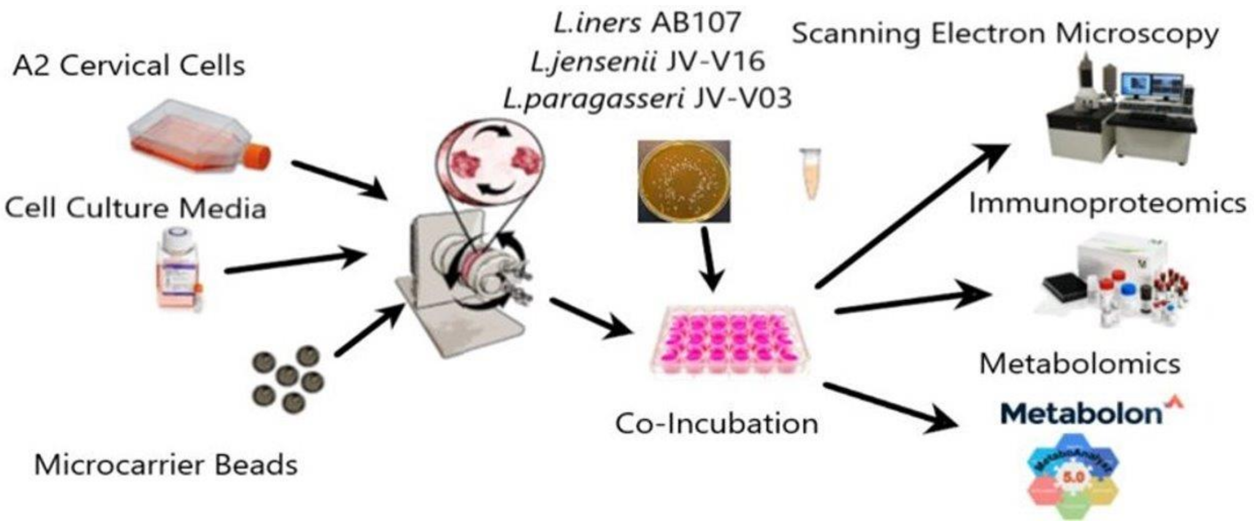
- **What is your study question?**
  - Does a particular species of bacteria associate with a diagnosis (cervical cancer)?
- **Will there be other factors you need to account for?**
  - Do the patients utilize cigarettes, douching hygiene, have history of BV? Are the patients pregnant or postmenopausal? Do we need to collect HPV status or genotypes?
- **How are you collecting your data?**
  - Clinical Team coordination at clinic, IRB consent, colposcopy biospy to derive diagnosis group
- **Will it be interpretable by computers?** -> Interoperable to others
  - Yes, and working through it for future use (two other studies have utilized already)
- How will you store your data?
  - Stored at NIH Sequence Read Archive and data use agreement through Women's Health Office due to IRB consent form. Code at Github.

From Experiment to Basic Science Metadata Management

How many of you work in wet lab/field research?

What type of metadata do you encounter?

From Experiment to Basic Science Metadata Management



[Jimenez et al, Msphere. 2023.]

Sample	Treatmen	(14 or 15)-	(S)-3-hydr	1,2-dioleo
PBS_116	PBS	405,887	76,009	206,658
PBS_118	PBS	847,138	108,050	177,241
PBS_120	PBS	511,835	58,643	274,290
PBS_123	PBS	667,936	64,681	259,518
PBS_124	PBS	392,964	106,190	239,745
PBS_125	PBS	485,556	59,092	110,821
L. paragas	L.paragass	700,587	94,376	321,208
L. paragas	L.paragass	985,133	75,839	207,689
L. paragas	L.paragass	660,800	90,472	410,015
L. jensenii	L.jensenii	527,266	67,944	398,375
L. jensenii	L.jensenii	381,775	90,894	205,589
L. jensenii	L.jensenii	463,027	78,480	219,223
L. jensenii	L.jensenii	247,194	88,932	350,906
L. iners_1	L.iners	506,928	98,945	267,280
L. iners_1	L.iners	492,555	56,428	138,259
L. iners_1	L.iners	591,220		840,044

From Recruitment to Clinic Metadata Management

How many of you work in clinical research?

What type of metadata do you encounter?

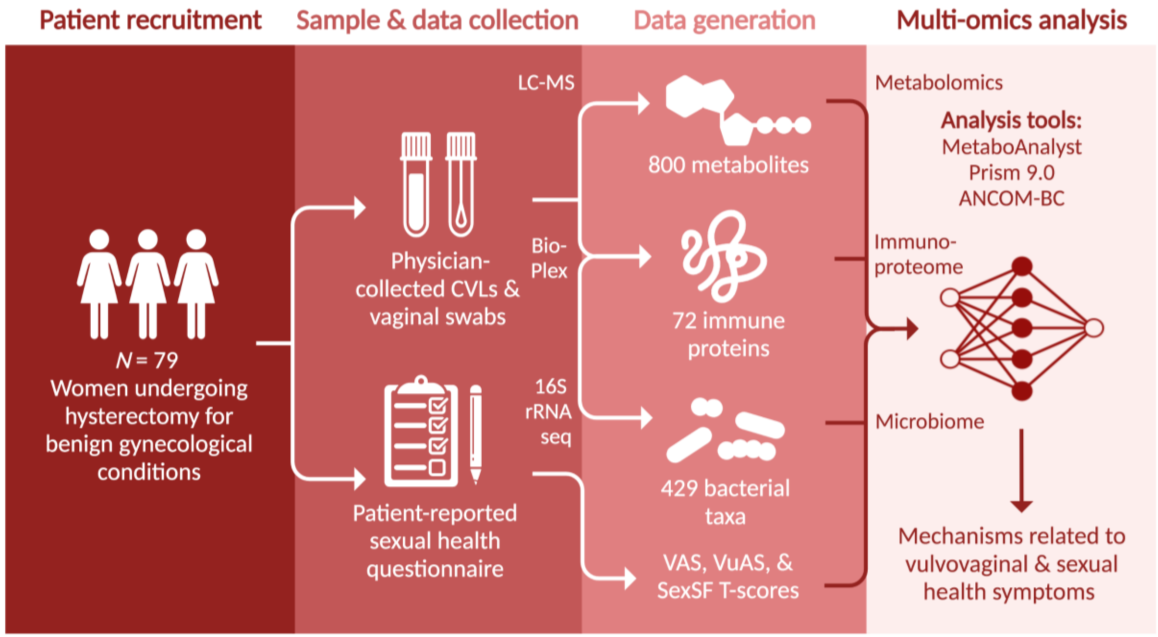
From Recruitment to Clinic Metadata Management

Medical/Clinic data:

Diagnosis for study: EMC, Hyperplasia, Benign condition

Benign condition diagnosis: fibroids, adenomyosis, endometriosis, abnormal uterine bleeding, etc

Grade, size, location, invasion, molecular tests



Crossley et al, unpublished, 2023

Laboratory Data:

HPLC metabolite concentrations

Microbiome 16S read counts

Milliplex – immune marker concentrations

Validated Survey Data:

Vaginal Assessment: EMC, Hyperplasia, Benign condition

Vulvar assessment: fibroids, adenomyosis, endometriosis, abnormal uterine bleeding, etc

Sexual health and Satisfaction questionnaire:

Clinical Data Importance and Data Stewardship

<u>Clinical Data and Information Security (NIH, but applicable elsewhere):</u>

- Proper handling of the vast domain of clinical data that is being continually generated from a range of data producers is a challenge for biomedical research community.

- Patient-related data is both quantitative and qualitative and can arise from a wide array of sources, including specialized research projects and trials; epidemiology; genomic analyses; clinical-care processes; imaging assessments; patient-reported outcomes; environmental-exposure records; and a host of social indicators now linked to health such as educational records, employment history, and genealogical records.

  - **<u>Key take-homes:</u>**

    - Practice robust and proactive information-security approaches to ensure appropriate stewardship of patient data

    - Curation of authentic, trusted data sources for future research

    - Protect against patient data compromise or loss.

https://www.niaid.nih.gov/research/clinical-metadata-standard

Clinical Data Importance

All of Us
RESEARCH PROGRAM

https://www.researchallofus.org/

https://www.gida-global.org/care

Electronic Health Record (EHR)

Medical/Clinic data:

Diagnosis for study: EMC, Hyperplasia, Benign condition

Benign condition diagnosis: fibroids, adenomyosis, endometriosis, abnormal uterine bleeding, etc

Grade, size, location, invasion, molecular tests

TABLE 1. SUGGESTED ELECTRONIC HEALTH RECORD CONTENTS BY PROVIDER

Data provider	*Kinds of data constituting the electronic health record*
Pharmacist	Medicine prescribed (linked to information about chemical composition, interactions, side effects, etc.)
	Dosage
Physician	Comments about the patient's observable physical condition at time of examination (coded)
	Diagnosis
	Treatment plan
Radiologist	Diagnosis (coded) after reading computed tomography scans, magnetic resonance imaging scans, X-rays, etc.
Raw test results	Electronic versions of image-based tests
	Unanalyzed content from monitoring devices
Clerical staff	Demographics (coded), such as sex, age, blood type, and family history
Nurse	Vital signs at time of examination (coded), including blood pressure and temperature
	Self-reported symptoms

https://www.liebertpub.com/doi/10.1089/big.2013.0023

https://library.ahima.org/doc?oid=106378#.ZBywGnbMK5c

Survey

Validated Survey Data:
Vaginal Assessment: EMC, Hyperplasia, Benign condition

Vulvar assessment: fibroids, adenomyosis, endometriosis, abnormal uterine bleeding, etc

Sexual health and Satisfaction questionnaire:

How would you set up this metadata and data?

In last month, XYZ?

Never	Almost Never	Sometimes	Fairly Often	Very Often
0	1	2	3	4
○	○	○	○	○

Series of questions then total score calculated based on survey requirements

Recording Metadata

- **README:** A README File is a text file located in a project-related folder that describes the contents and structure of the folder and/or a dataset so that a researcher can locate the information they need.

- **Data Dictionary:** Also known as a *codebook*, a data dictionary defines and describes the elements of a dataset so that it can be understood and used at a later date.

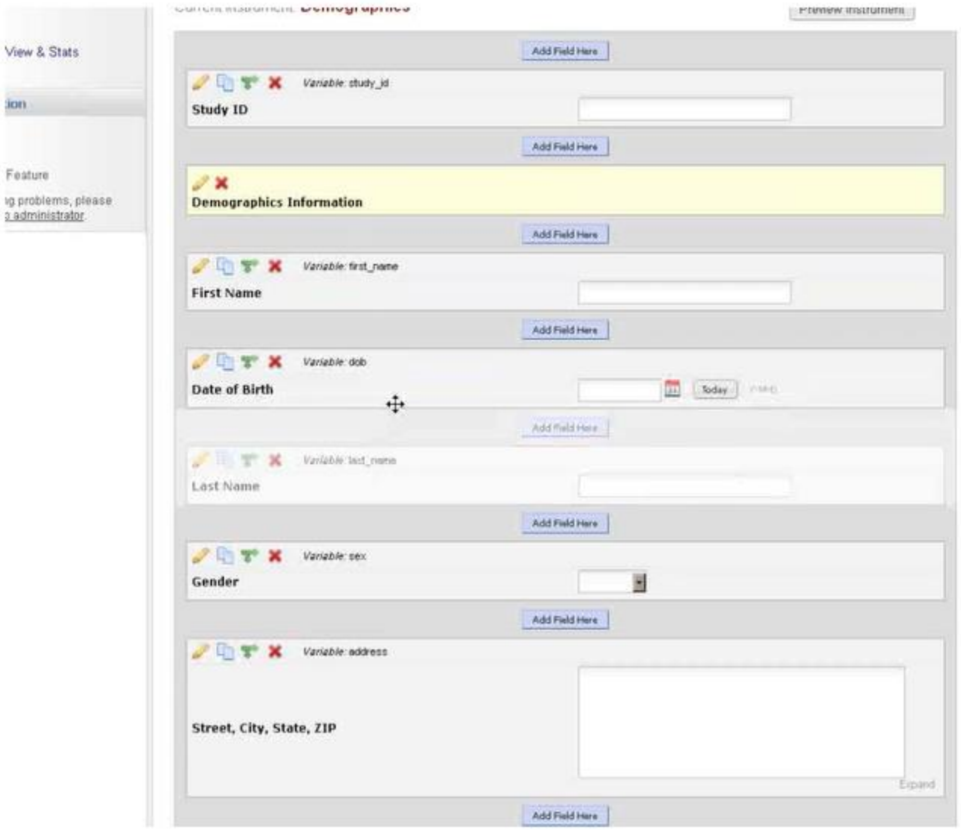
- **Protocol:** A protocol describes the procedure(s) or method(s) used in the implementation of a research project or experiment. If you need to maintain protocols, we strongly recommend a tool like protocols.io.

- **Lab Notebook:** For research groups that use them, Electronic Lab Notebooks offer several advantages over traditional paper notebooks.

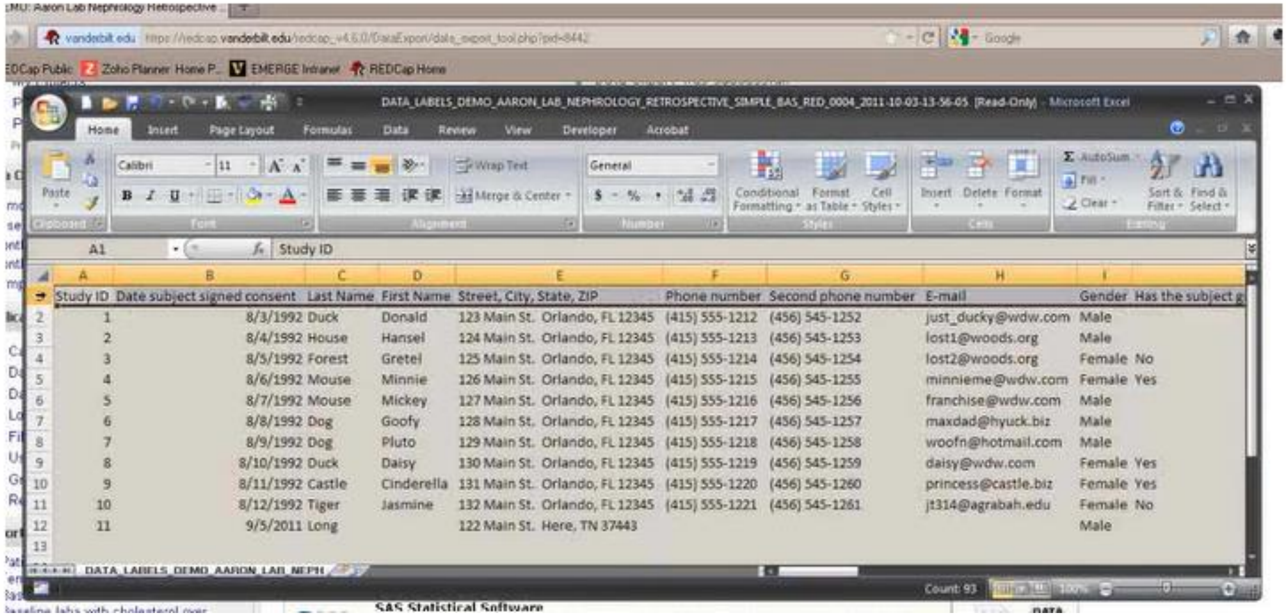
How do you currently record metadata?

https://datamanagement.hms.harvard.edu/collect-analyze/documentation-metadata

Planning ahead with ReDCaP



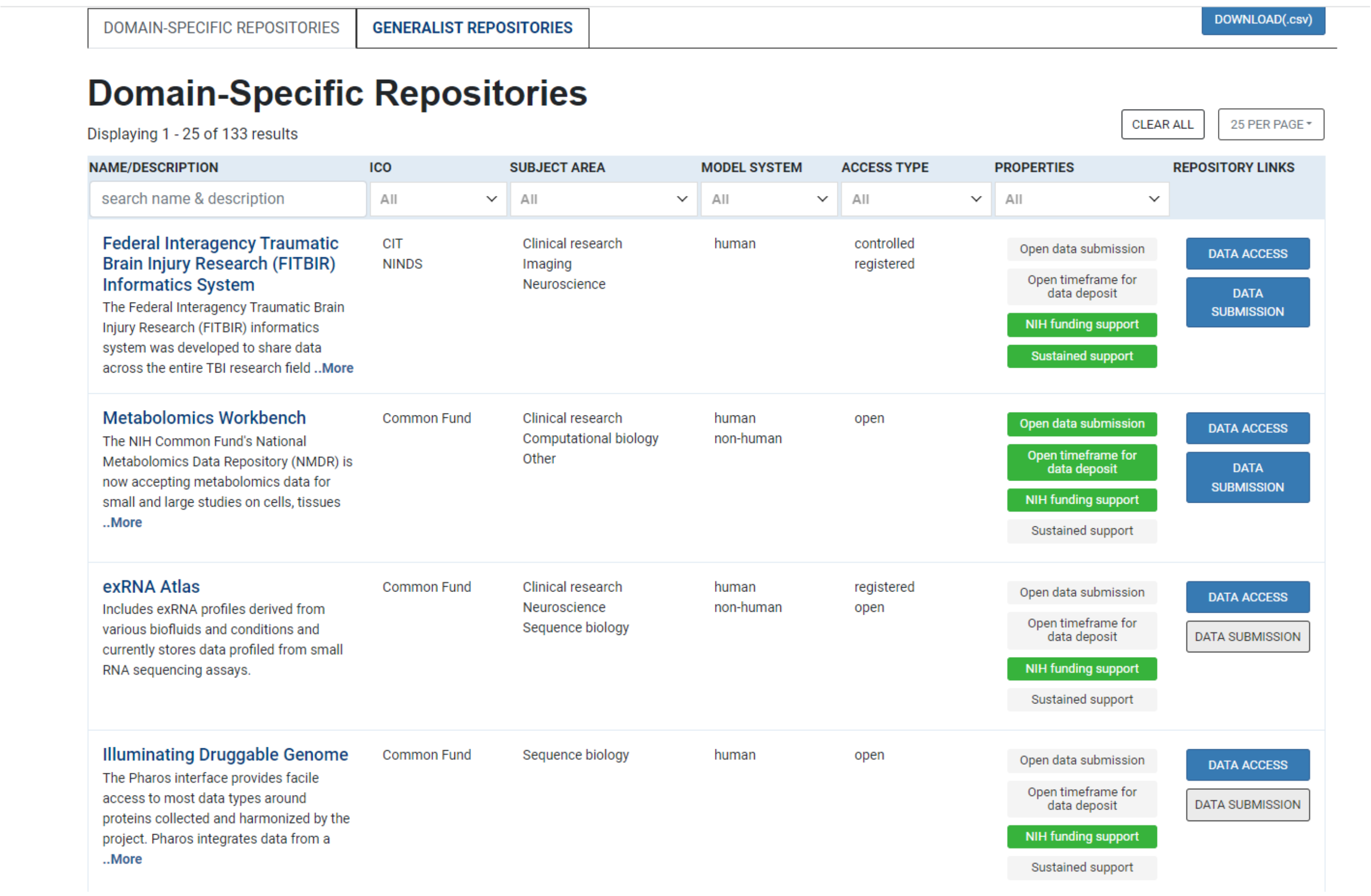
The University of Arizona Center for Biomedical Informatics and Biostatistics (CB2) is pleased to be a partner in the REDCap Consortium. It supports a secure web application (REDCap) designed exclusively to support data capture for research studies and institutional projects. **The REDCap application allows users to build and manage online surveys and databases quickly and securely.**

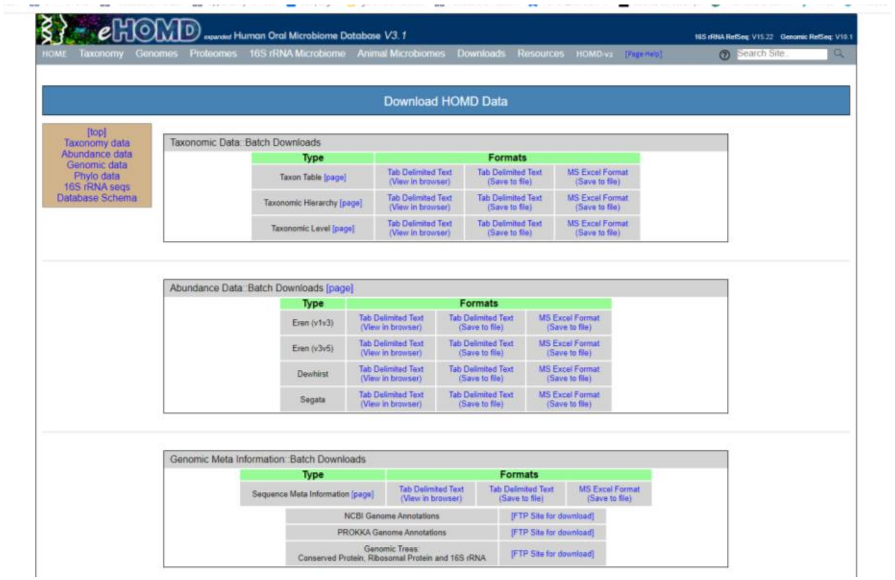


https://cb2.uahs.arizona.edu/services-tools/surveys-clinical-databases-redcap/redcap-training

https://cb2.uahs.arizona.edu/services-tools/surveys-clinical-databases-redcap

NIH Repositories





https://homd.org/

https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html

Public Databases are not Created Equal

- Data Repository submission: What you put in will be what you get out
  - Nicole's example as a comparative genomics Grad student:

Identifiers	BioSample: SAMD00290000; SRA: DRS179169
Organism	Bifidobacterium bifidum cellular organisms; Bacteria; Terrabacteria group; Actinomycetota; Actinomycetes; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium
Package	MIGS: cultured bacteria/archaea; version 6.0
Attributes	

sample name	B. bifidum BI-28
collection date	2017-02-17
broad-scale environmental context	missing
local-scale environmental context	human-associated habitat
environmental medium	feces
geographic location	Japan
host	Homo sapiens
isolation and growth condition	doi.org/10.1038/s41396-021-00937-7
latitude and longitude	missing
number of replicons	1
project name	Infant gut SCFA and microbiota relationship
reference for biomaterial	missing
strain	BI-28

Metagenome or environmental sample from human gut metagenome

Identifiers	BioSample: SAMN27962527; Sample name: PRL2010-MUC-C; SRA: SRS12856508
Organism	human gut metagenome unclassified entries; unclassified sequences; metagenomes; organismal metagenomes
Package	Metagenome or environmental; version 1.0
Attributes	

host	not applicable
isolation source	not applicable
collection date	2022-03-08
geographic location	Italy
latitude and longitude	not collected
source_organism	Bifidobacterium bifidum

BioProject	PRJNA833139 Retrieve all samples from this project

https://www.nature.com/articles/sdata201921

Public Databases are not Created Equal

- Data Repository submission: What you put in will be what you get out

Figure 3: Metadata submissions to NCBI BioSample from 2009–2017.

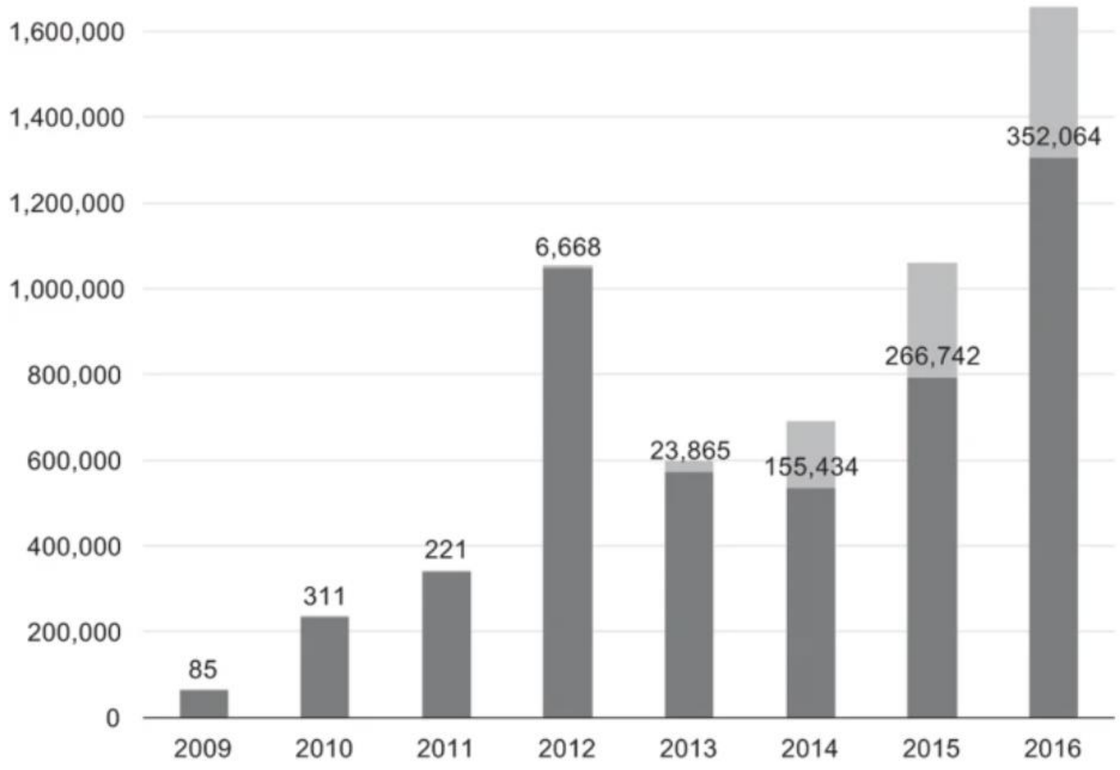
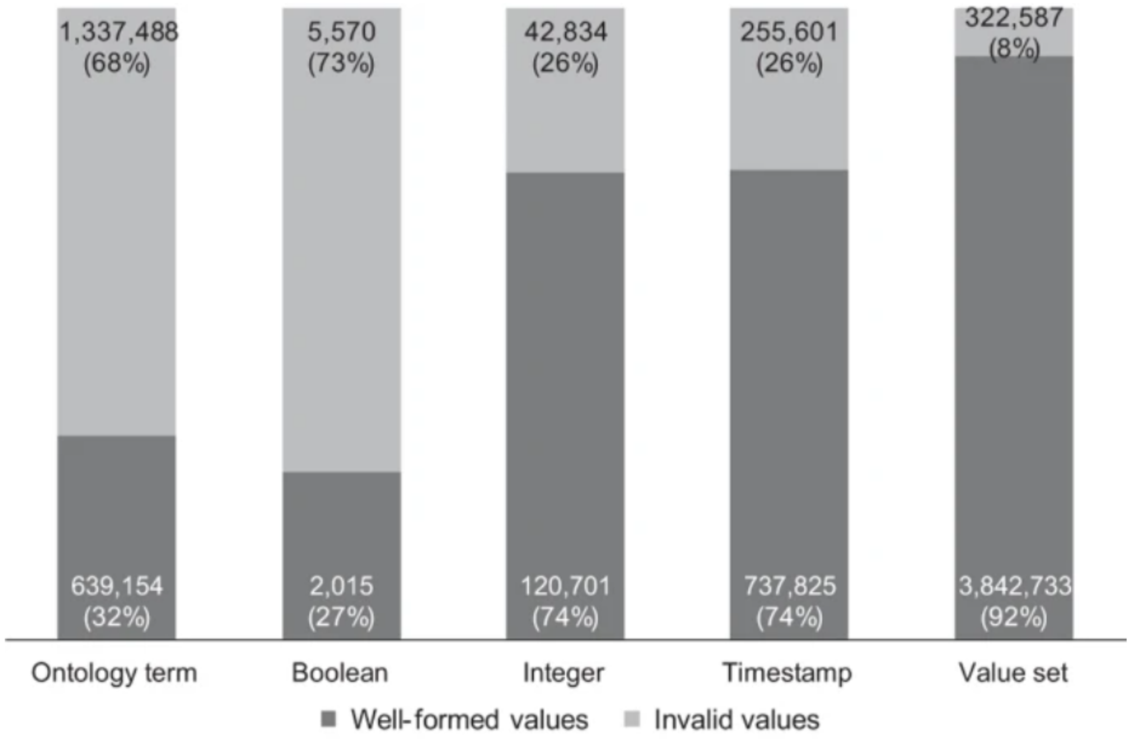


Figure 4: Quality of dictionary attributes in NCBI BioSample according to their type.



https://www.nature.com/articles/sdata201921

Public Database Inspection

Considerations:

- Does this dataset fit my needs? answer my research question?
- Do I trust the data from this dataset? What are vairables or metadata that might key me in on the validity?
- What is the format? Data type?
- Is it in a format that I can I run basic statistics or run exploratory analyses to investigate the data set? Is the data complete?

There are many resources here at UA and on GitHub to help you inspect your data, Check out the calendar and office hours of DSI to figure out best ways to investigate your data!

https://realpython.com/pandas-python-explore-dataset/ https://link.springer.com/chapter/10.1007/978-3-030-69150-9_2

Preparation of Data/Metadata

Data also exist in a wide variety of formats, which complicates the ability of researchers to find and use biomedical research data generated by others and creates the need for extensive data "cleaning."

According to a 2016 survey, data scientists across a wide array of fields said they spend most of their **work time (about 80 percent) doing what they least like to do: collecting existing datasets and organizing data**.

That leaves less than **20 percent of their time for creative tasks like mining data for patterns** that lead to new research discoveries.

NIH Data Science Strategic Plan

Data Cleaning

Box 1. Terms Related to Data Cleaning

Data cleaning: Process of detecting, diagnosing, and editing faulty data.

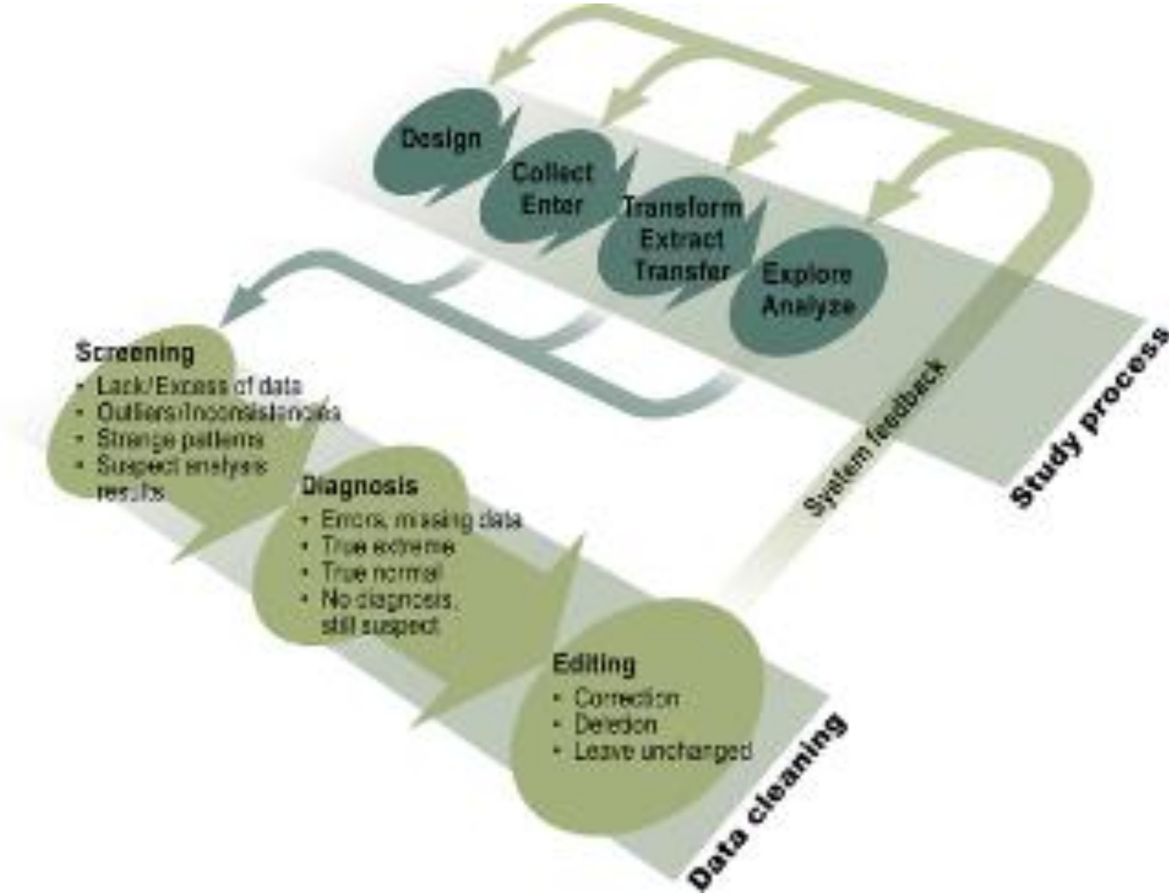
Data editing: Changing the value of data shown to be incorrect.

Data flow: Passage of recorded information through successive information carriers.

Inlier: Data value falling within the expected range.

Outlier: Data value falling outside the expected range.

Robust estimation: Estimation of statistical parameters, using methods that are less sensitive to the effect of outliers than more conventional methods.



https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020267

http://varianceexplained.org/r/tidy-genomics/

https://careerfoundry.com/en/blog/data-analytics/what-is-data-cleaning/

Formatting for Analysis

Excel is easy to use, flexible and powerful, however, it often gives us too much freedom which leads to bad practices and difficult to re-use data and metadata.

Items to Keep in the Computer's Mind:

- Using multiple tables in a sheet
- Using multiple tabs
- Not filling in true zeros
- Using problematic null values
- Using formatting(highlighting, bolding, merging columns) to convey information
- Placing comments or units in cells
- Entering more than one piece of information in a cell
- Inconsistency in used values
- Using problematic field names(beware the space)
- Using special characters in data
- Inspect Data name for misnumbering or dates
- Etc....

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

https://carpentries-incubator.github.io/fair-bio-practice/07-data-in-excel/index.html

Formatting for Analysis

Check your File type and your software dependencies!!!

- Plain text files like comma or tab separated values (.csv, .tsv) can be accessed without any special software.
- If you analyse your data with R or Python, or you know that your data are meant to be processed that way you should be using text formats whenever possible, and as soon as you capture your data.
- if you only use Excel and so does your community, just keep using it. Just be aware of the possible pitfalls discussed, especially when working with gene or protein names and accession numbers.

If utilizing a Graphical User Interface or specialized software, check the sample data files this will save time and help you figure out what formats may be required if not explicitly stated

https://carpentries-incubator.github.io/fair-bio-practice/07-data-in-excel/index.html

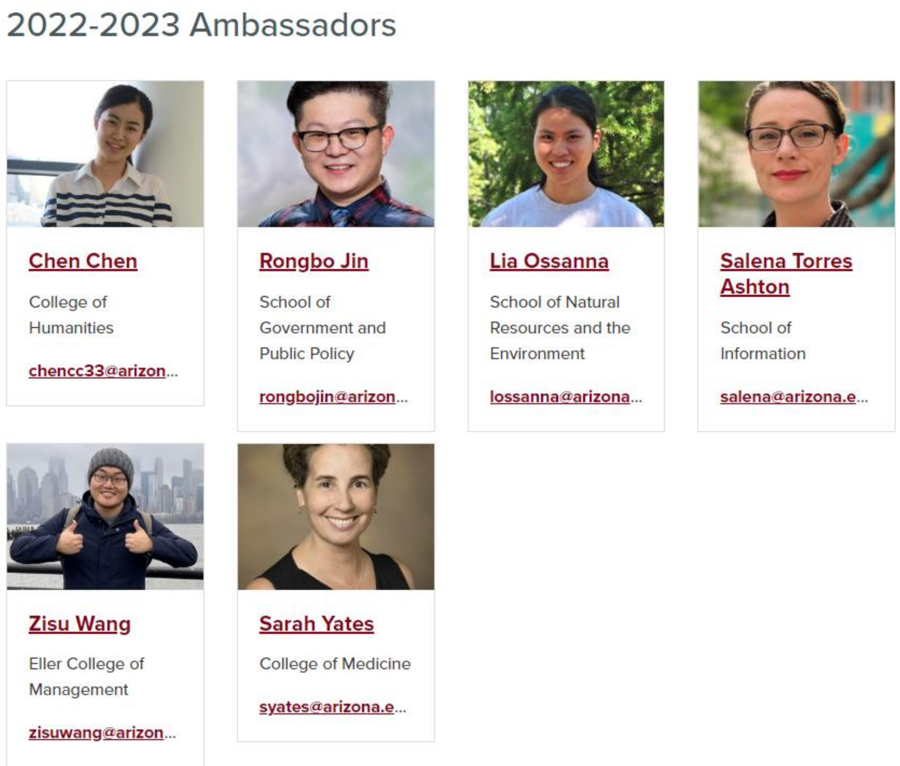
Overview Metadata? - Exercise

- What contextual details (metadata) are needed to make your data meaningful?
- What form or format will the metadata describing your data take? Which metadata standards will you use? If there is no applicable standard, how will you describe your data in a way that will make them accessible to others?
- How will metadata files be generated for each of the data sets that you produce? Who will do the work of data description?
- Who on your team will be responsible for ensuring that metadata standards are followed and are correctly applied to the corresponding data sets?

https://guides.lib.umich.edu/c.php?g=283277&p=1888478

Data Science Fellowship or Ambassador Programs

Each semester, a new cohort featuring up to 12 fellows will be offered. Fellows are expected to attend and participate in twice weekly virtual training activities. **Fall 2023 semester** participants will be awarded a $7,000 stipend for successful completion of the program.



In short, DSAs are champions for data science literacy in your college!

https://datascience.arizona.edu/education/data-science-fellows

https://datascience.arizona.edu/dsa

Data Science on Phoenix Campus

Viewing/Panel discussion in the works!


THE UNIVERSITY OF ARIZONA RESEARCH, DISCOVERY & INNOVATION
Data Science Institute

Workshop - Data Science on the Phoenix Biomedical campus!

Let's talk Data Science on the UA Phoenix Biomedical Campus!

- Interested in Open Science or Data Science?
- Want to know how to improve the reproducibility of your research and speed up your data analysis workflow?
- Concerned about the new NIH mandate and Data Management Plans?

REGISTER FOR THIS WORKSHOP

November 9th and December 7th
3:30-5 p.m.

HSEB room A446, UA Health Sciences Library – Phoenix. Enter via the library main entrance on the 3rd floor and then take the stairs or the elevator within the library to reach room A446.

Or join virtually with the zoom link that you will receive upon registration.

YOU ARE NOT ALONE!

Join other like-minded researchers and Luisa Rojas, a Data Science Institute Data Science Fellow and CTS doctorate student, as she hosts a series of engaging workshops dedicated to open science principles, Data Management Plans and Open Data Science tools for the lab – right here at the UA Phoenix Biomedical Campus.

Bring your questions and suggestions and, of course, feel free to bring a friend!

VentureCafe: Hackathon
April 6th 11:30 am to 5 pm
PBC information

Future Programming focused on DS Initiatives

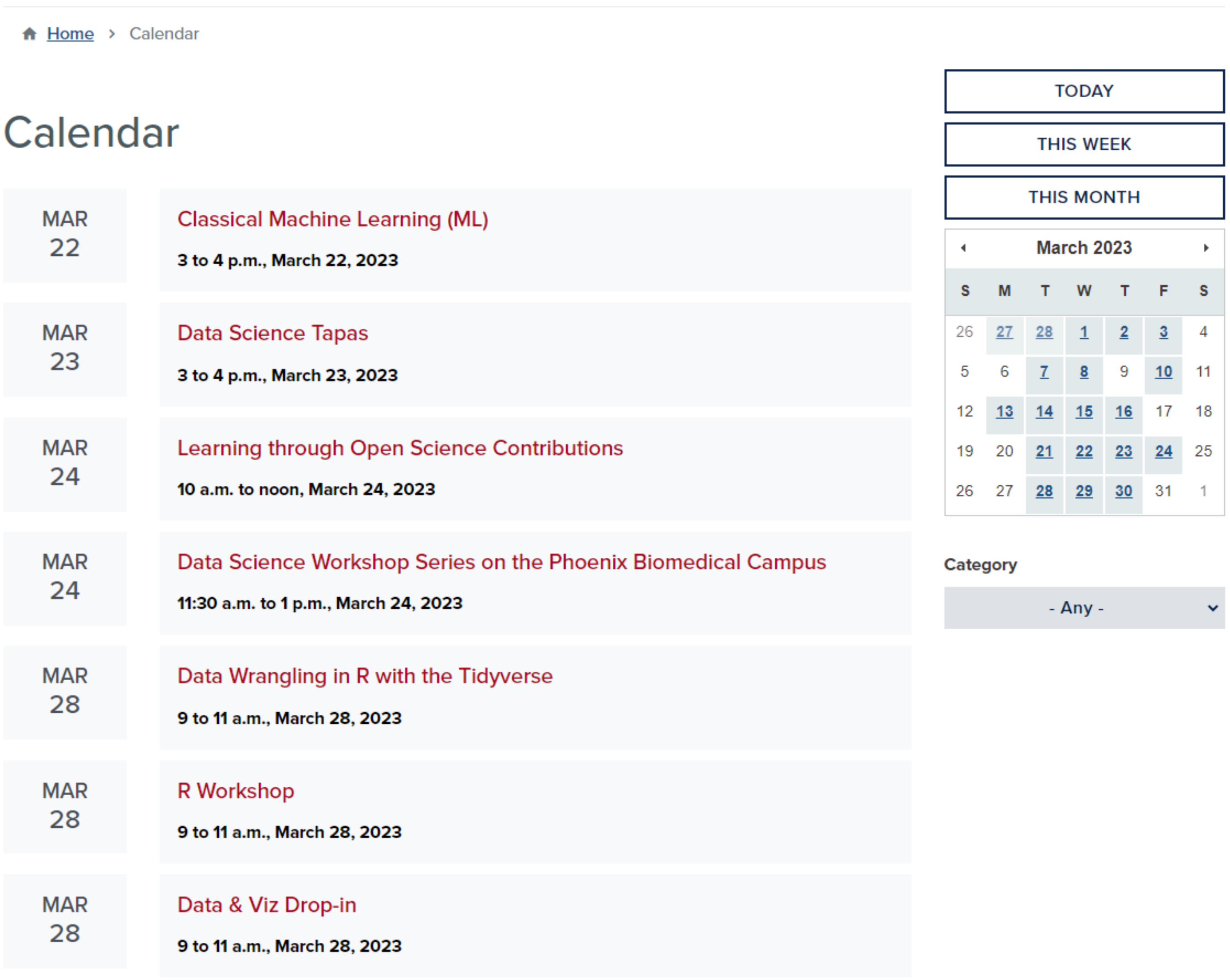
1. *Computational and Experimental Research*
2. *Clinical and Translational Research*
3. *Epidemiology and Population-Based Research*
4. *Research Using Data Generated Outside a Research Context*

Booz Allen identified five competencies necessary for making data FAIR and AI-ready:

Competency	Description
Dataset Documentation	• Documentation of study design and data elements, including methodology, study population description, inclusion/exclusion criteria, sampling procedures, and expected relationships between variables through use of conceptual modeling or knowledge graphs. • Ethical, legal, and social implications of poor or incomplete dataset documentation on the analysis of current studies and potential secondary usage.
Ontology Usage & Data Encoding	• Domain-specific standardized ontologies and when to request the addition of new content, attachment of granular and accurate metadata to datasets through use of secure data/metadata entry software (such as REDCap). • Self-identification in human-subject data collection (i.e., free response of race/ethnicity, sex/gender, and disability status) and compliance with federal data collection standards).
Data Cleaning & Formatting	• Importing of structured and unstructured data into a coding environment, removal of identifiable information from datasets before sharing, and transformation of data into a structured, machine-readable format that is cleaned, accurate, and consistent. • Ethics of handling self-identification in human-subject data collection including 'roll-up' to federal classification standards and interpretation of classification categories in findings, as well as an understanding that these are evolving concepts that may not be accurately/consistently reflected in older datasets.
Data Governance	• Creation of long-term management plan for data, including curation, sharing, access, reuse, and archiving/preservation. • Human-subject data storage and sharing standards, including the need to de-identify or perturb data to remove PII or PHI prior to sharing or release; and a sufficient understanding of study context and design to identify and document potential threats to inference.
Data Collaboration	• Facilitation of interdisciplinary collaboration across biomedical and data science research teams during iterations of study design, data collection, and model development/analysis, which requires sufficient knowledge of AI concepts to assist biomedical researchers with algorithmic data needs and sufficient knowledge of biomedical concepts to convey to AI developers relevant and precise biomedical contextual information, to the degree that biomedical and AI researchers can generate and assess clinically meaningful results. • Avoidance of or mitigation for damaging ethical implications in deployed health models, such as (a) poor data collection and aggregation methods, namely misappropriation of data collected in non-research efforts without consideration of intended collection context and/or misapplication to target populations without accounting for differences between study and target populations; (b) inadequately documented and/or labeled historical or discriminatory bias in input data, and (c) insufficiently rigorous explorations of the effects that findings from algorithmic deployments could have on patients if incorporated into clinical workflows, including the difference between statistically and clinically significant findings and future model retraining due to increasing noncomparability between pre- and post-treatment datasets.

We want to hear from you, what do want to see on the PBC!

Data Science Institute Calendar



Need help with your data?

Check out Data Science Institue Calendar, there is likely a workshop or office hours to assist in your analysis journey!

https://datascience.arizona.edu/calendar

Data Science Resources at University of Arizona

Foundational Open Science Skills

Open Science Framework

Roots for Resillience

Data Management

Soteria

Cyverse

ReData

Learn to Code

rstudio-connect

PlanetMicrobe

University of Arizona - High Perfomance Computing

UA Data Science Slack Channel

ResBaz

Women in Data Science - Tucson

Acknowledgements

RESEARCH, INNOVATION & IMPACT
Data Science Institute

- **Funding**
  - Data Science Institute – Data Science Fellows
  - BIO5 Institute – Postdoctoral Fellowship
  - Community Foundation for Sothern Arizona

- **<u>Herbst-Kralovetz Lab Members and Clinical Team</u>**
- **College of Medicine-Phoenix - Department of Obstetrics and Gynecology**
- **Early Investigator Support**
  - Postdoctoral Affairs Office
  - UA-CoM-P Research Office,
  - Native American Cancer Prevention – GUIDeS: U54CA143924, U54CA143925

References/resources

- https://www.getty.edu/publications/intrometadata/setting-the-stage/
- https://guides.lib.umich.edu/c.php?g=283277&p=1888478
- https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/
- https://www.ncbi.nlm.nih.gov/grc
- https://metadata-wizard-tutorial.readthedocs.io/en/latest/
- https://databrowser.researchallofus.org/
- https://datacarpentry.org/
- https://cran.r-project.org/web/packages/eatGADS/vignettes/meta_data.html
- https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
- https://allofus.nih.gov/
- https://hdsr.mitpress.mit.edu/pub/gg6swfqh/release/2

Hands-On:

There are many ways to clean/inspect data:

- Python, R, etc

- For today's example we will utilize a tool recommended by Data Carpentries called OpenRefine: https://openrefine.org/
- Activity: https://datacarpentry.org/OpenRefine-ecology-lesson/
- Follow the tutorial or try it out with your data

OpenRefine