

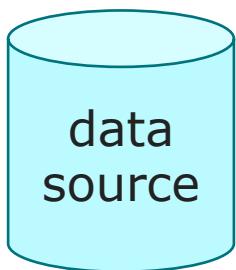
Working with Research Data

Sawyer Newman

Data Librarian for the Health Sciences

Working with Research Data

1. Collecting
2. Finding
3. Organizing
4. Wrangling
5. Analysis
6. Documentation
7. Sharing



data
source



Selecting



Storage, Organization & Security

Collecting Data

Data might come from...

- Instruments
 - Microscopes
 - Surveys
 - RNA sequencing machines
- Electronic health records
- Web scraping
- Online databases

Clinical data collection software

Qualtrics

- yalesurvey.qualtrics.com

RedCap

- redcapportal.yale.edu

Oncore

- medicine.yale.edu/ycci/oncore

Data collection groups at Yale

- Core Research Facilities
 - Instrumentation
 - Project design
 - Training

research.yale.edu/core-research

Finding Data

Finding data from literature

Supplemental Material

Files stored and made available with full-text article

Data Availability Statement

Text description in full text (may include a citation for the data)

Data Citations

Machine-readable metadata in references or article text

Filtering for data in PMC

The screenshot shows the NCBI PMC search interface. At the top, there's a navigation bar with the NCBI logo, 'Resources' (with a dropdown arrow), and 'How To' (with a dropdown arrow). Below the navigation is the PMC logo and text: 'US National Library of Medicine' and 'National Institutes of Health'. The search bar contains the query: 'PMC ((brain[Title]) AND brain[Title]) AND cushing'. Below the search bar are links for 'Create alert', 'Journal List', and 'Advanced'. On the left side, under 'Article attributes', there's a section titled 'Associated Data' which is checked (indicated by a blue circle with a white checkmark). A callout box with a teal border and a teal arrow points to this section with the text: 'Select the Associated Data filter under "Article attributes"'. Other filter options listed under 'Article attributes' include 'Author manuscripts', 'Digitized back issues', 'MEDLINE journals', 'Open access', and 'Retracted'. Below these are sections for 'Text availability' ('Include embargoed articles'), 'Publication date' ('1 year', '5 years', '10 years', 'Custom range...'), and 'Download Tools'. The main search results area has a heading 'Search results' and a sub-heading 'Items: 1 to 20 of 28'. It includes a note: 'Filters activated: Associated Data. [Clear all](#) to show 281 items.' The first result is a list item: '1. Asmae Belhaj, Laurence Dewachter, Sandrine Rorive, Myri Melot, Emeline Hupkens, Céline Dewachter, Jacques Crete Benoît Rondelet PLoS One. 2017; 12(7): e0181899. Published online 2017 Jul 28. doi PMCID: PMC5533440 Article PubReader PDF–16M Citation'.

PMC Data Box

Associated Data

▼ Supplementary Materials

S1 Data: Datasets used and analyzed in the current study. Brain death (BD) (CO), systemic arterial pressure (SAP), right atrial pressure (RAP), Noradrenaline (NA), epinephrine (EPI), arterial partial pressure of oxygen (PaO₂), arterial partial pressure of carbon dioxide (PaCO₂), arterial oxygen saturation (SpO₂), pulmonary artery occlusion pressure (PAWP), pulmonary artery pressure (PAP), pulmonary vascular resistance (PVR), arterial PO₂ divided by PaO₂ (PCP), venous compartmental resistance (Rv), arterial PO₂ divided by arterial partial pressure of oxygen (PaO₂/FiO₂), acute lung injury score (ALI Score), anti-myeloperoxidase immunoglobulin G (mPOD IgG), interleukin 6 (IL-6), interleukin 8 (IL-8), interleukin 1β (IL-1β), interleukin 10 (IL-10), interleukin 2 (IL-2), tumor necrosis factor alpha (TNF-α), Bcl2 associated athanoptosis (Bak) protein 1 (Bak1), B-cell lymphoma 2 (Bcl-2), intercellular adhesion molecule 1 (ICAM-1), vascular cell adhesion molecule 1 (VCAM-1), hypoxia-inducible factor 1α (HIF-1α), Glutathione peroxidase 4 (GPX4), glutathione peroxidase 6 (GPX6), hypoxia-inducible factor 1α (HIF-1α), Glutathion peroxidase response 1 (OXSR-1).

(XLSX)

[pone.0181899.s001.xlsx](#) (27K)

GUID: C3576FE8-A66F-4BEB-902D-B0596C5B4721

▼ Data Availability Statement

All relevant data are within the paper and its Supporting Information file.

Finding other data

- NIH Data Sharing Repositories
- Cushing/Whitney Medical Library Data Quick Search tool
- Ask the library!

Library Data Quick Search

If you would like to add an existing published dataset added to this list please, email medicaldata@yale.edu.

Search

- Any -

SEARCH

RESET

DATA SOURCE	DESCRIPTION	TOPICS	AFFILIATED INSTITUTIONS
-------------	-------------	--------	-------------------------

CDC Data Catalog

Datasets and data visualizations from the Centers for Disease Control and Prevention.	<ul style="list-style-type: none">• Administrative data• Biomonitoring• Disability & health & toxicology• Injury• Vaccination• Violence• Pregnancy• Disability & health	Centers for Disease Control and Prevention (CDC)
---	--	--

Epic Data Requests (JDAT)

The Joint Data Analytics Team (JDAT) at Yale provides customized reporting and data analysis from the Epic data system.	<ul style="list-style-type: none">• Electronic health record data
---	---

Epic, Yale University

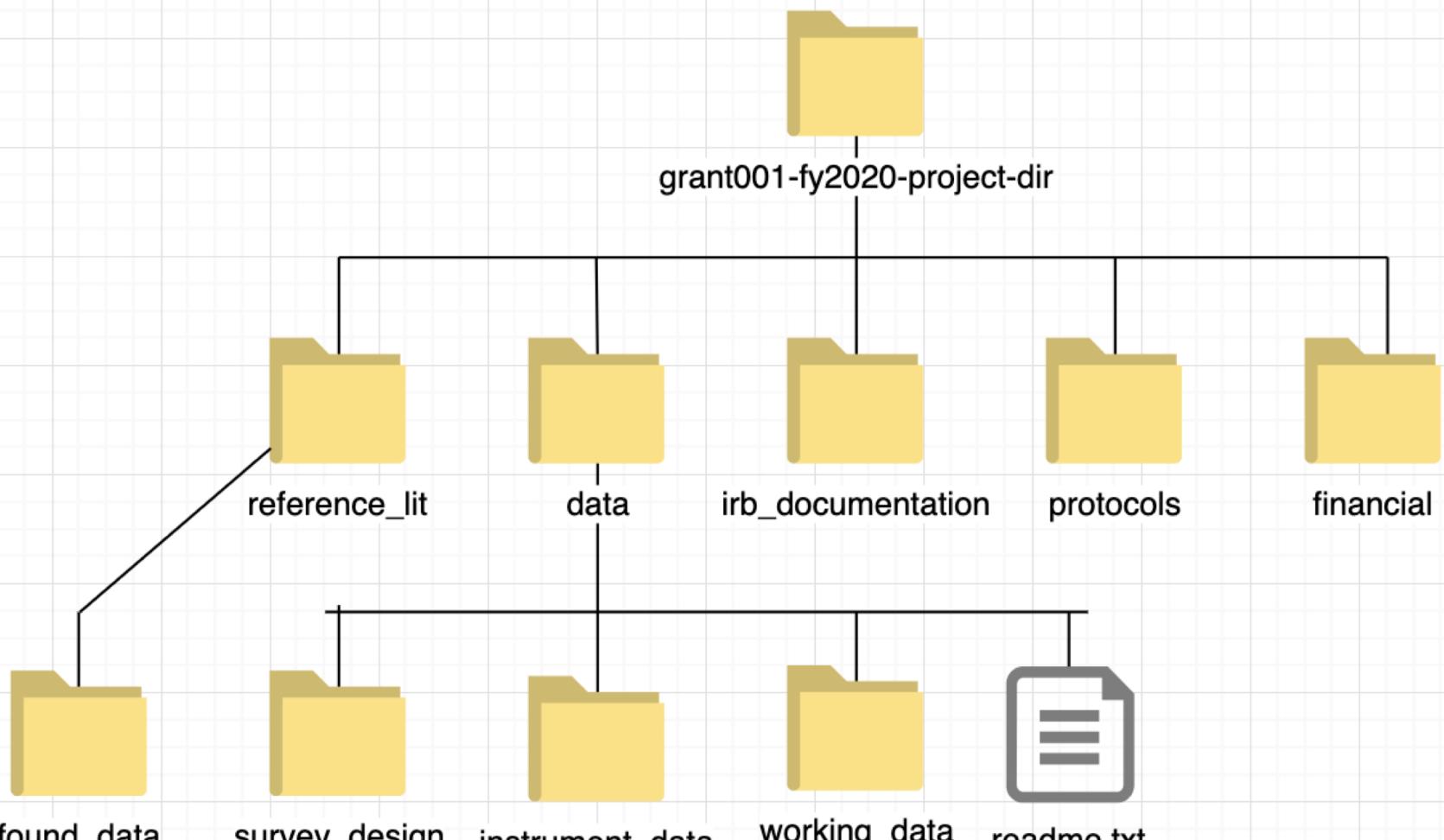
Organizing Data

File & folder naming

- Create unique and simple files names
- Use only alpha-numeric characters. Avoid using special characters such as: ? / \$ % & ^ # . \ : < >
- Use underscores (_) and dashes (-) to represent spaces
- Use leading zeros with the numbers 0-9 to facilitate proper sorting
- Dates should follow the ISO 8601 standard of YYYY_MM_DD or YYYYMMDD

Folder organization

- Use a consistent organization template per project, per analysis, etc.
- Create documentation detailing organization rules
- Create wide directories, not deep directories



Storage

Working Data

- Secure box
- Spinup at Yale
- Storage@Yale Active and Object Tiers

Archived Data

- Storage@Yale's Archive Tier

Research Data Documentation

Documentation

Information you note about your data and research

Documentation may make note of:

- Data variable meanings
- File organization structures
- Methodologies employed
- Context of your data and research
- Project contributors

Forms of documentation

- ReadMe Files
- Data Dictionaries
- Codebooks
- Version control

Wrangling Data

Stages of Data Wrangling

1. Selection: data matches project scope
2. Cleaning and standardization: data is consistent in meaning and format
- 3. Reshaping: data is presented in a useful way**
4. Aggregating: data summaries

Wide vs Tall Data

id	2014-value	2015-value	2016-value
1	101	103	106
2	206	254	244
3	394	369	311

id	year	value
1	2014	101
1	2015	103
1	2016	106
2	2014	206
2	2015	254
2	2016	244
3	2014	394
3	2015	369
3	2016	311

United vs Split Data

respondent	question-1-multi
Smith, Jane	a, b, d
Doe, John	b, c

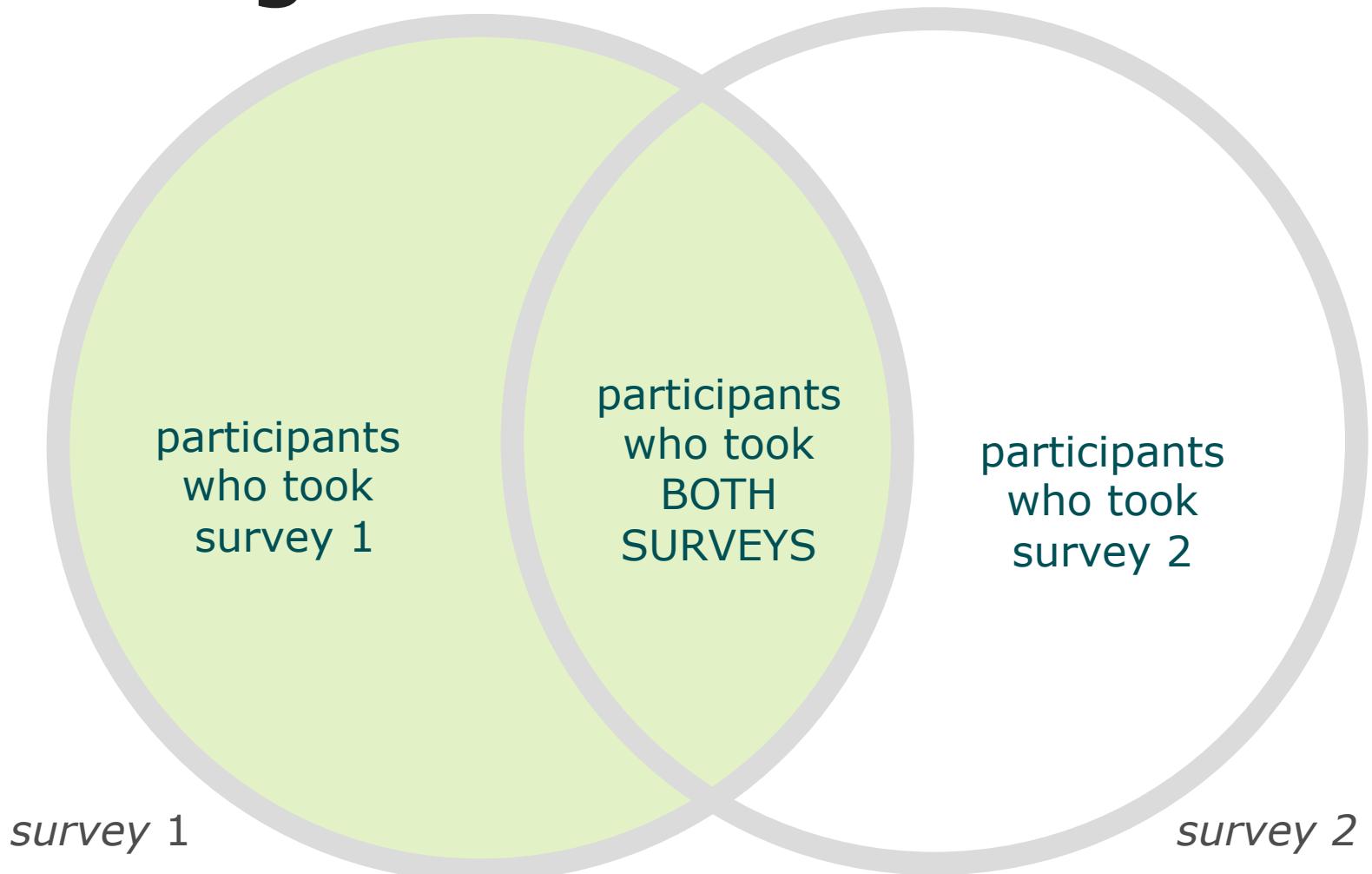
last-name	first-name	questi on-1- a	questi on-1- b	questi on-1- c	questi on-1- d
Smith	Jane	a	b	NA	d
Doe	John	NA	b	c	NA

Hierarchical vs Tabular Data

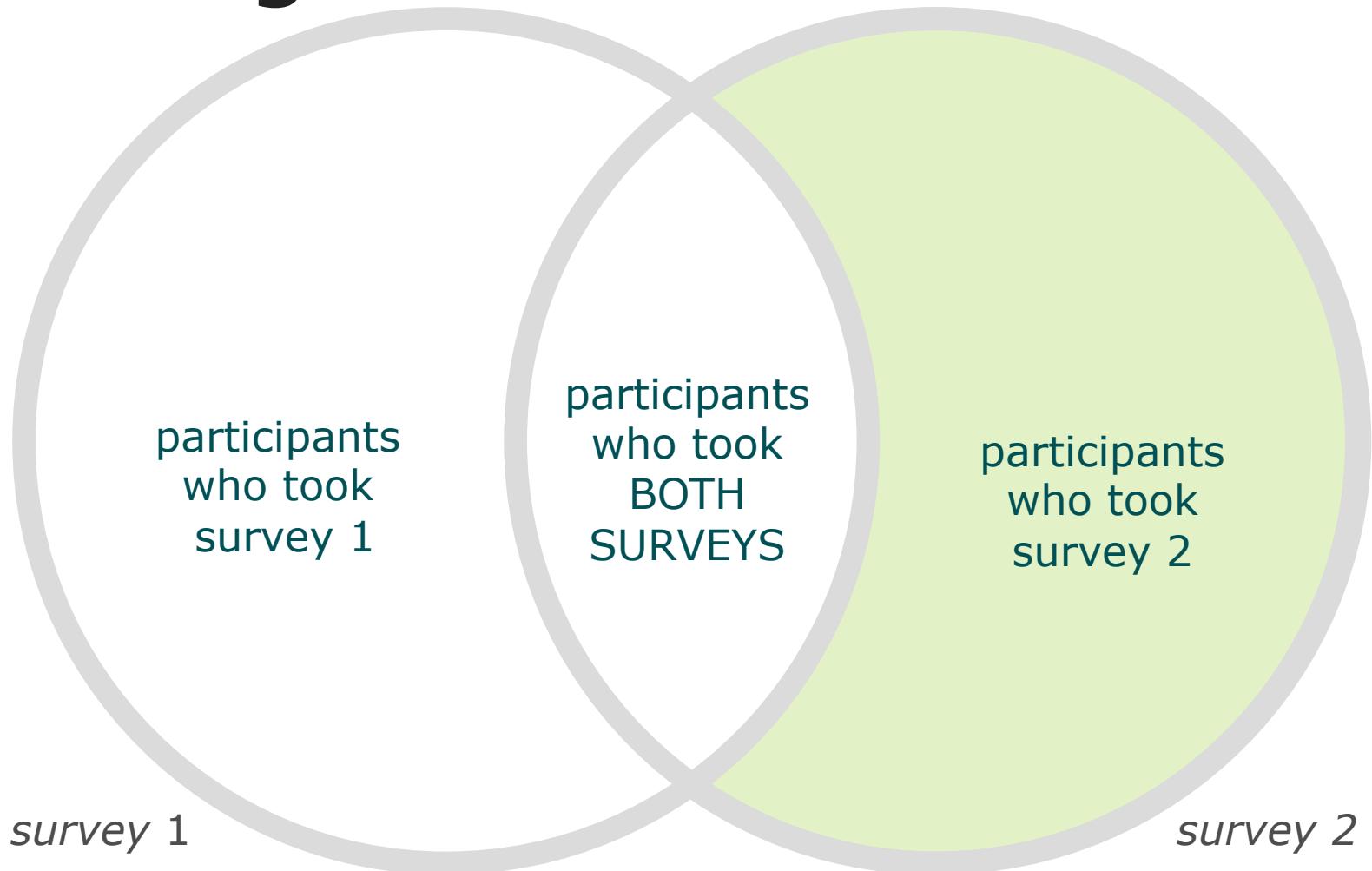
```
{  
  "samples": [  
    {  
      "id": "s001",  
      "test_id": "t254",  
      "creation_date": "2019-04-21"  
    },  
    {  
      "id": "s002",  
      "test_id": "t254",  
      "creation_date": "2019-04-22"  
    }  
  ]  
}
```

id	test_id	creation_date
s001	t254	2019-04-21
s002	t254	2019-04-22

Join Logic Visualized



Join Logic Visualized



Analysis

Analysis software

- NVIVO
- R, Python
- SAS, SPSS, STATA
- Excel

Find Data Analysis Help Services at Yale

- Yale Center for Analytical Sciences (YCAS)
- Yale Center for Clinical Investigation (YCCI)
- Yale Center for Research Computing (YCRC)
- Cushing/Whitney Medical Library Bioinformatics Support
- StatLab Services

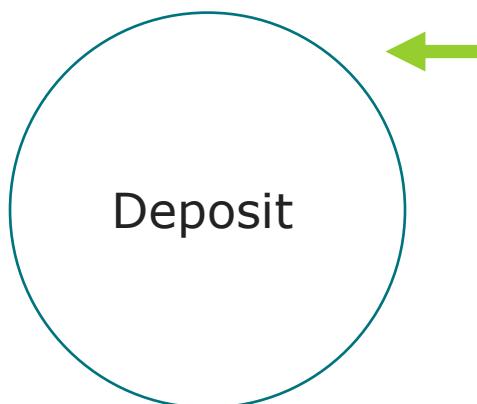
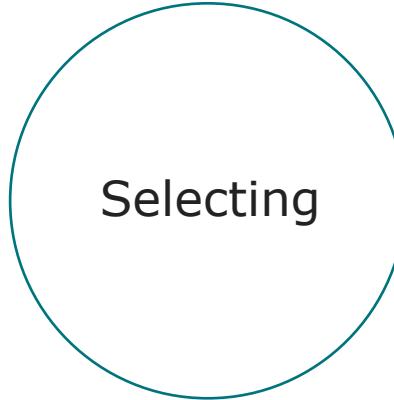
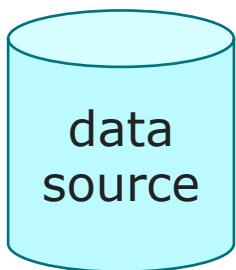
Data Sharing

Sharing Working Data

- Box and Box Secure
 - yale.account.box.com
- Yale Secure File Transfer
 - your.yale.edu/yale-link/secure-file-transfer

Data Depositing

- May be a requirement of your funder or publisher
- Choosing a data repository
 - By funding agency
 - By discipline
 - Cross-disciplinary repositories
- Depositing your data allows you to enhance your publication
- Data can include code



Storage, Organization & Security

Reach out with any questions



Data Management Plans



Data Tools & Software



Data Policy Guidance



Find Datasets



Data Storage



Best Practices & Definitions



Data Support Groups at Yale



Consultations & Drop-Ins

medicaldata@yale.edu