Working with Research Data

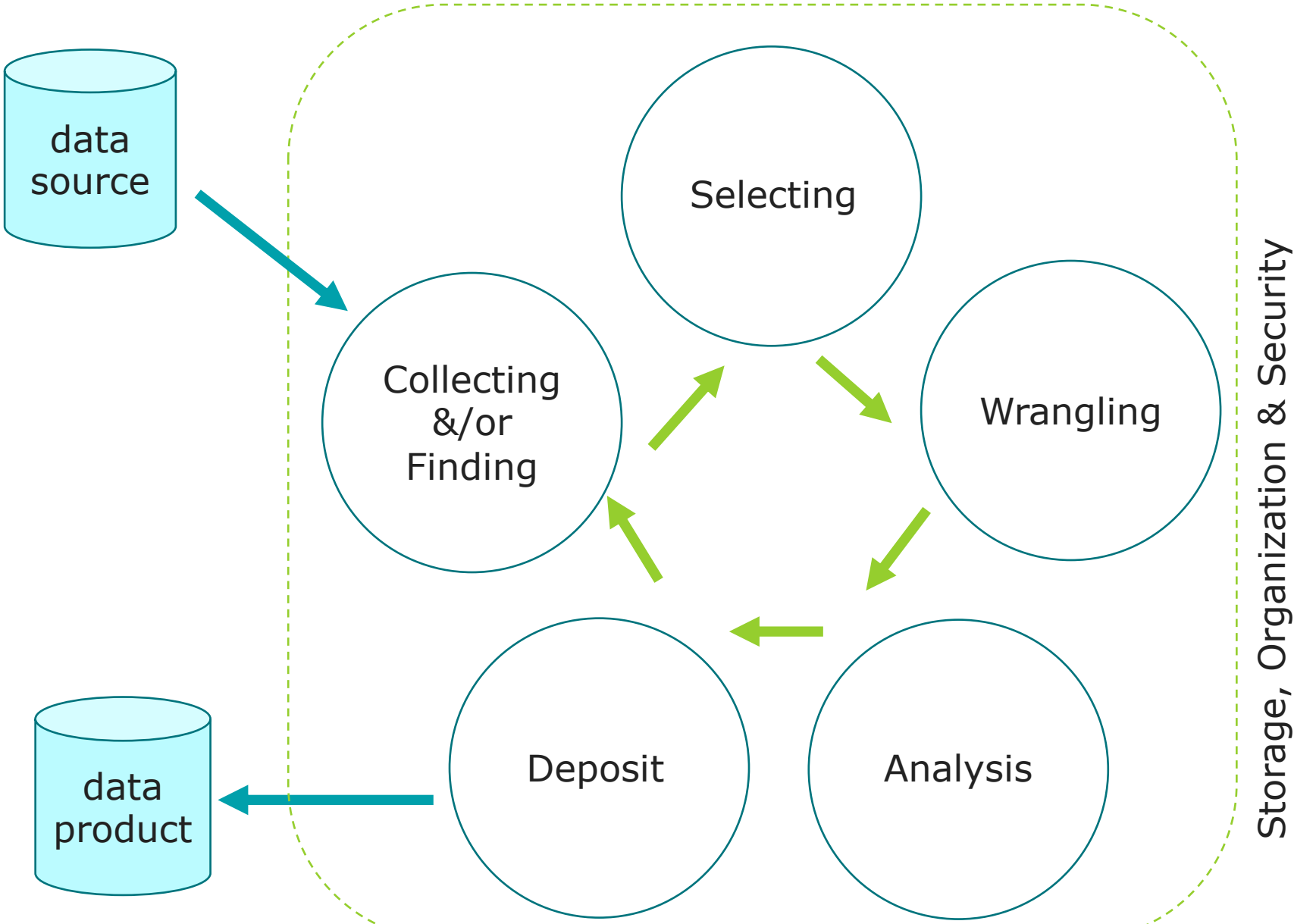
Sawyer Newman

Data Librarian for the Health Sciences

Yale *Harvey Cushing / John Hay Whitney Medical Library*

Working with Research Data

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


data source
Selecting
Collecting &/or Finding
Wrangling
Deposit
Analysis
data product
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

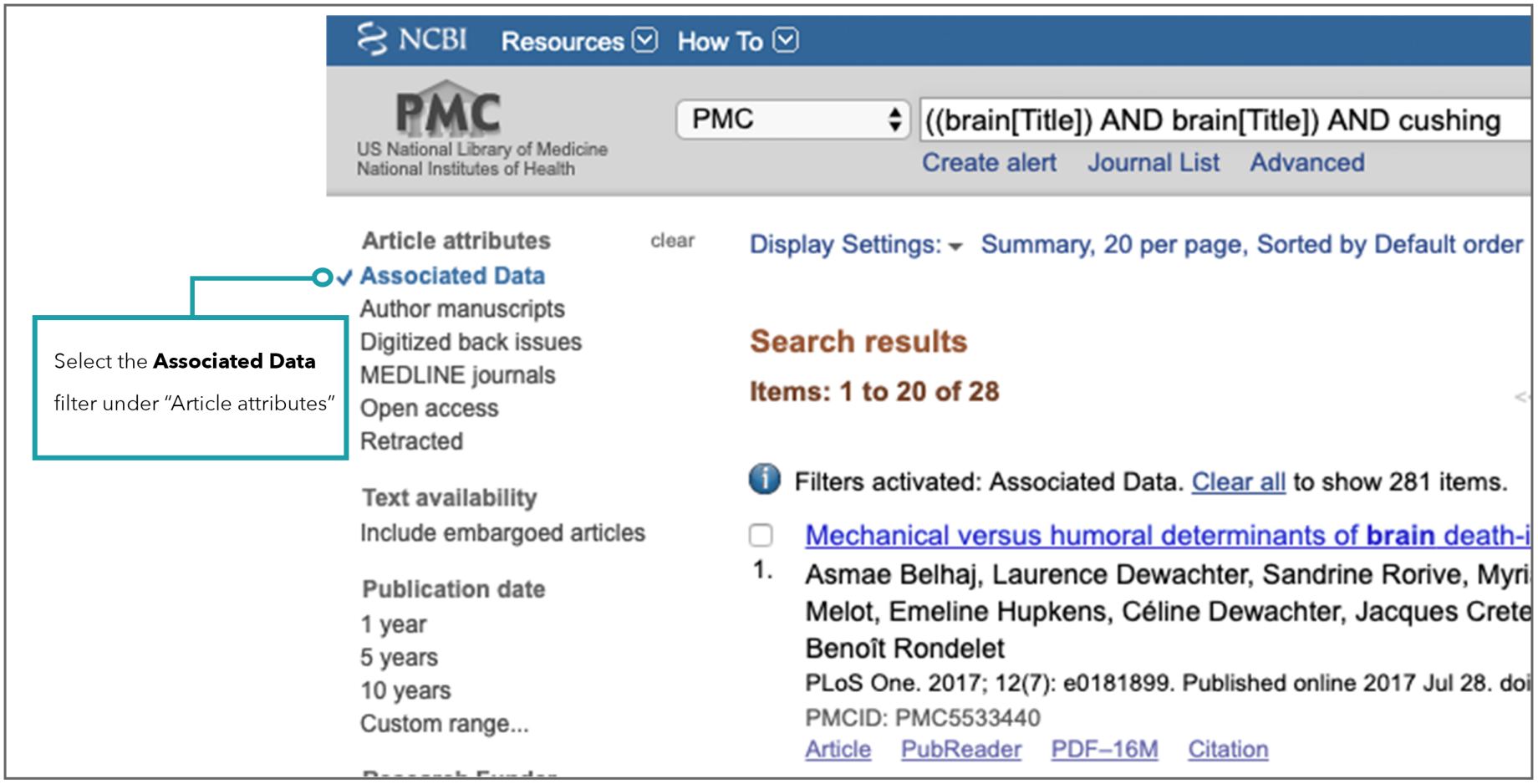
Data Availability Statement

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

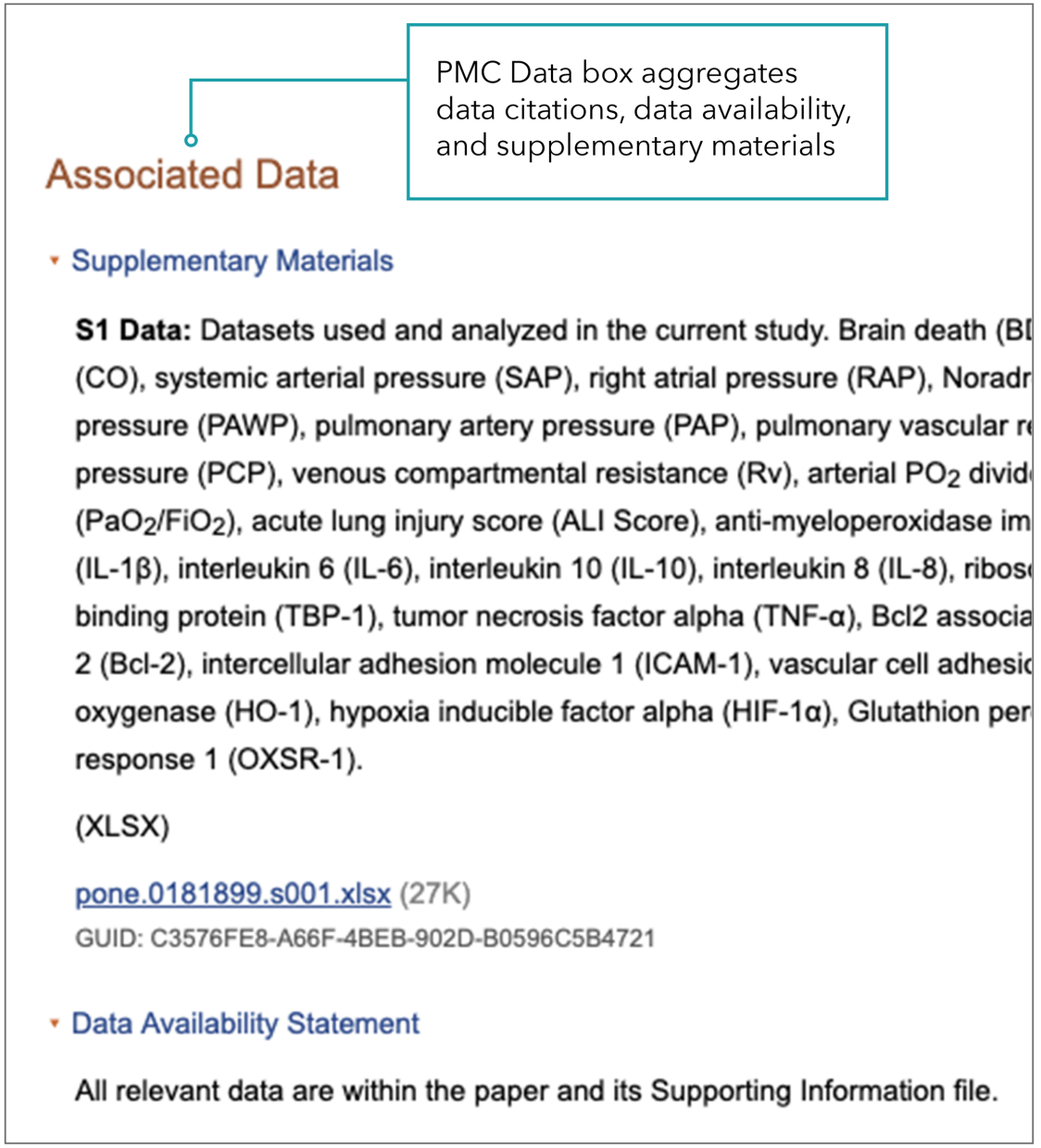
Data Citations

Machine-readable metadata in references or article text

Filtering for data in PMC



PMC Data Box


PMC Data box aggregates data citations, data availability, and supplementary materials
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), Noradr
pressure (PAWP), pulmonary artery pressure (PAP), pulmonary vascular r
pressure (PCP), venous compartmental resistance (Rv), arterial PO_2 divid
(PaO_2/FiO_2), acute lung injury score (ALI Score), anti-myeloperoxidase im
(IL-1β), interleukin 6 (IL-6), interleukin 10 (IL-10), interleukin 8 (IL-8), ribos
binding protein (TBP-1), tumor necrosis factor alpha (TNF-α), Bcl2 associa
2 (Bcl-2), intercellular adhesion molecule 1 (ICAM-1), vascular cell adhesi
oxygenase (HO-1), hypoxia inducible factor alpha (HIF-1α), Glutathion per
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

Affiliated Institutions

- Any -

SEARCH RESET

DATA SOURCE	DESCRIPTION	TOPICS	AFFILIATED INSTITUTIONS
CDC Data Catalog	Datasets and data visualizations from the Centers for Disease Control and Prevention.	• 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.	• Electronic health record data	Epic, Yale University

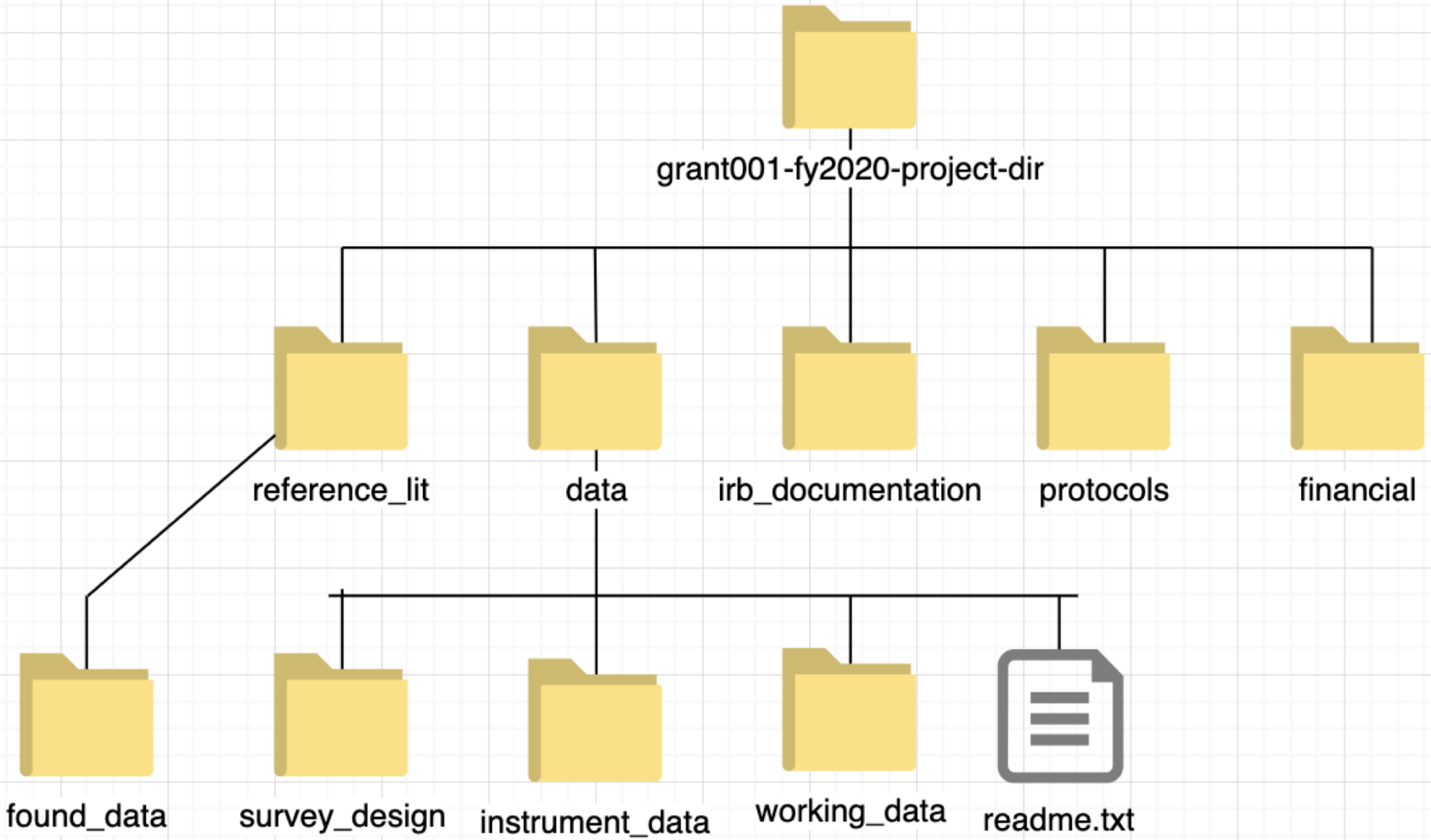
Organizing Data

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


grant001-fy2020-project-dir
reference_lit
data
irb_documentation
protocols
financial
found_data
survey_design
instrument_data
working_data
readme.txt

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

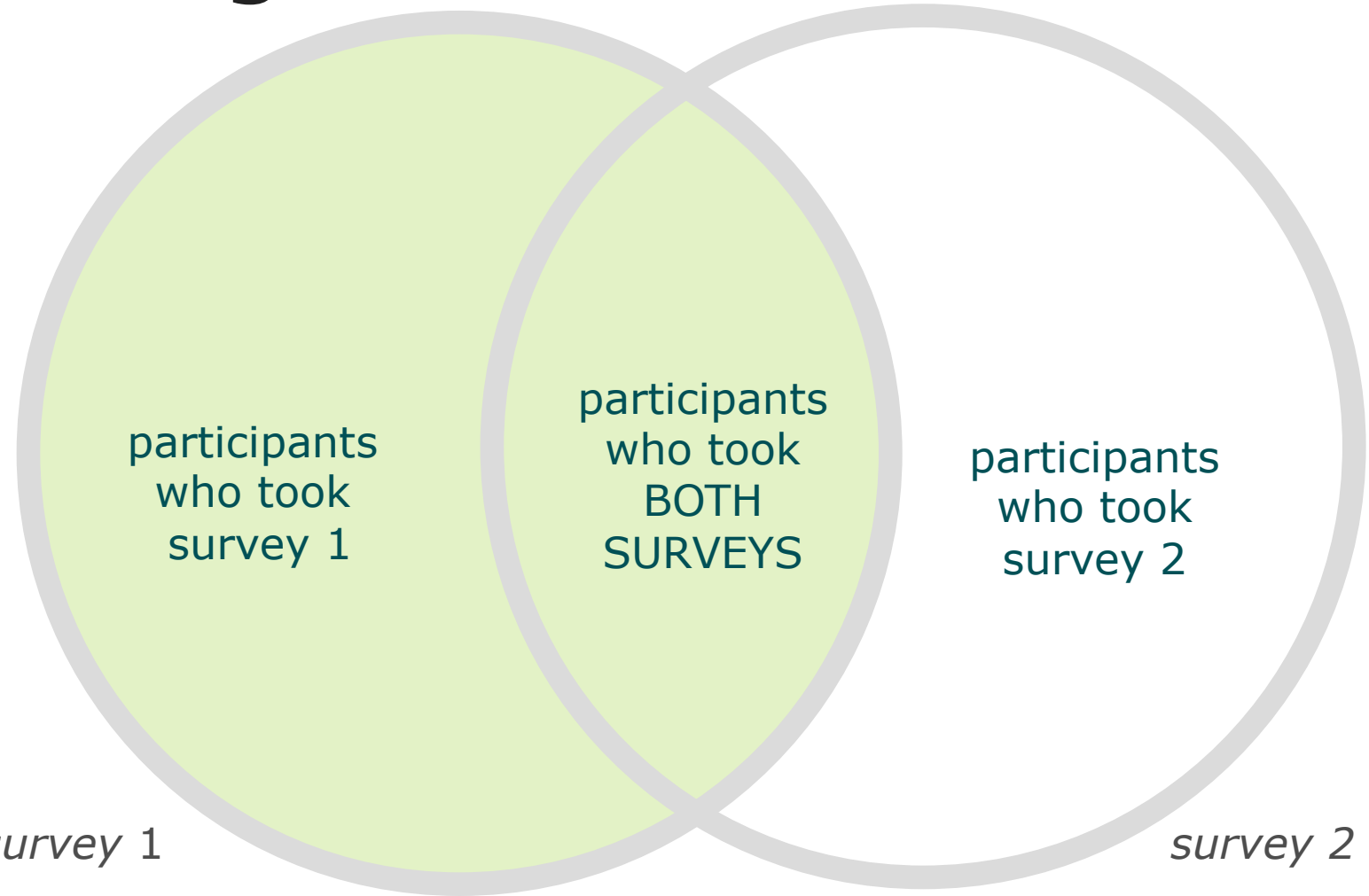
last-name	first-name	question-1-a	question-1-b	question-1-c	question-1-d
Smith	Jane	a	b	NA	d
Doe	John	NA	b	c	NA

Hierarchical vs Tabular Data

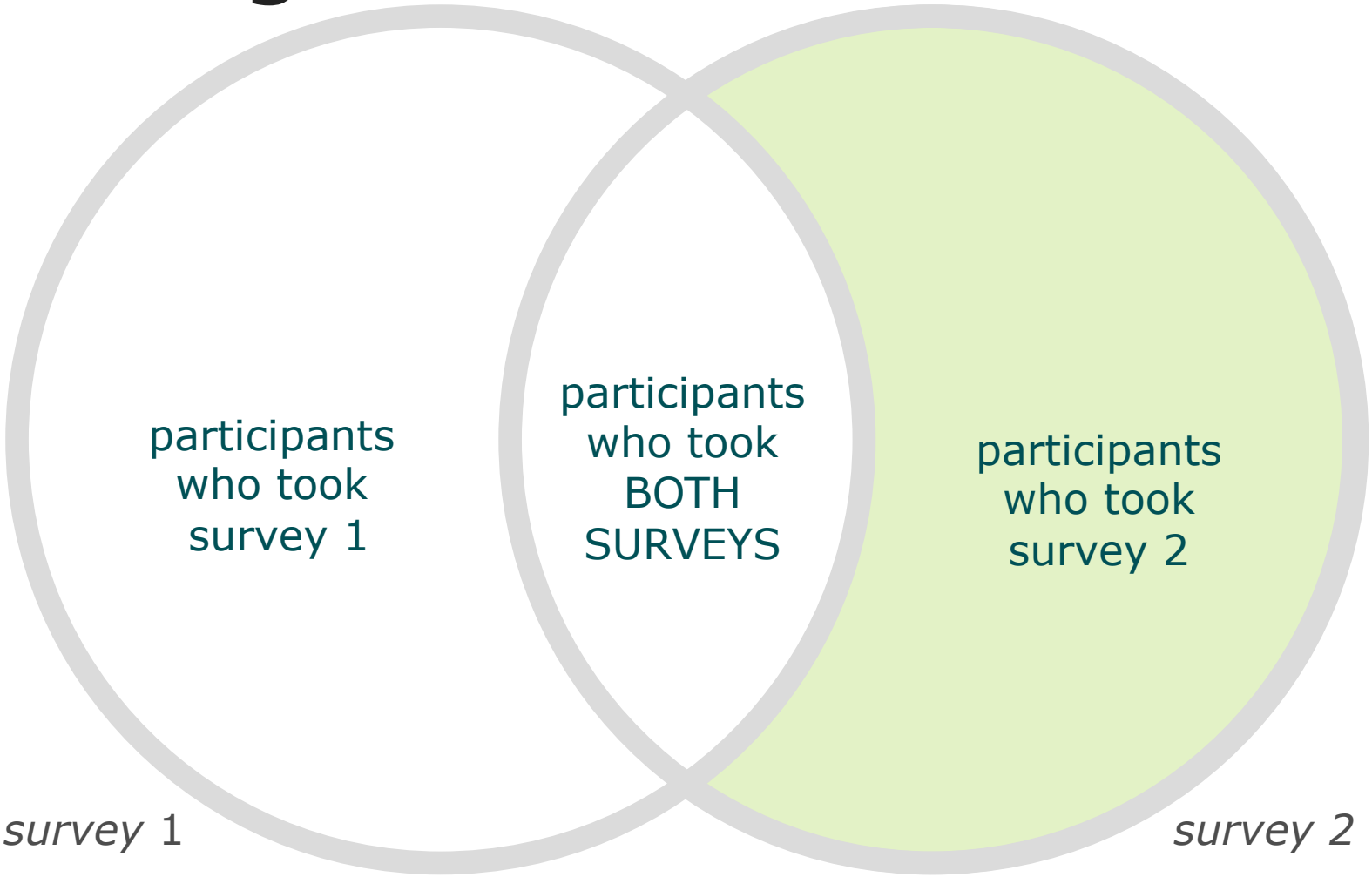
```
{
”samples”: [
{
“id”: “s001”,
“test_id”: “t254”,
”creation_date”: “2019-04-21”
},
{
“id”: “s002”,
“test_id”: “t254”,
”creation_date”: “2019-04-22”
}
]
}
```

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

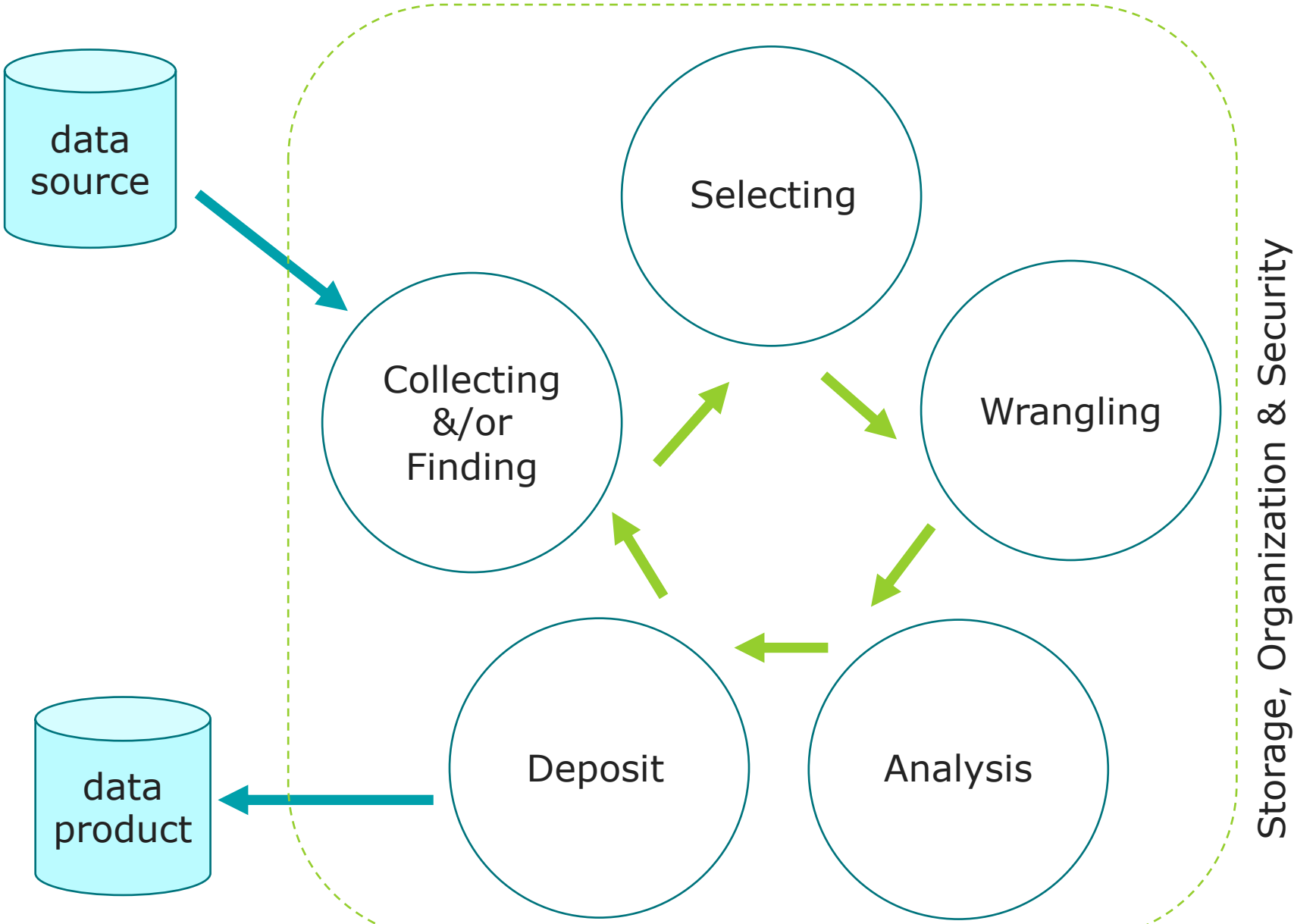
Data Sharing

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


data source
Selecting
Collecting &/or Finding
Wrangling
Deposit
Analysis
data product
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