# Core Concepts of Working with Data

Yale UNIVERSITY LIBRARY
*Harvey Cushing/John Hay Whitney Medical Library*

# What considerations should you make when working with data?

- Data format (file or spreadsheet organization)
- Data sources
- Number of files
- File size
- File format
- Computational need
- Data use agreements (DUAs) and licenses
- Data security
- Data sharing requirements

# What considerations should you make when working with data?

- Data format (file or spreadsheet organization)
- Data sources

Data Literacy

- Number of files
- File size
- File format
- Computational need

Technological

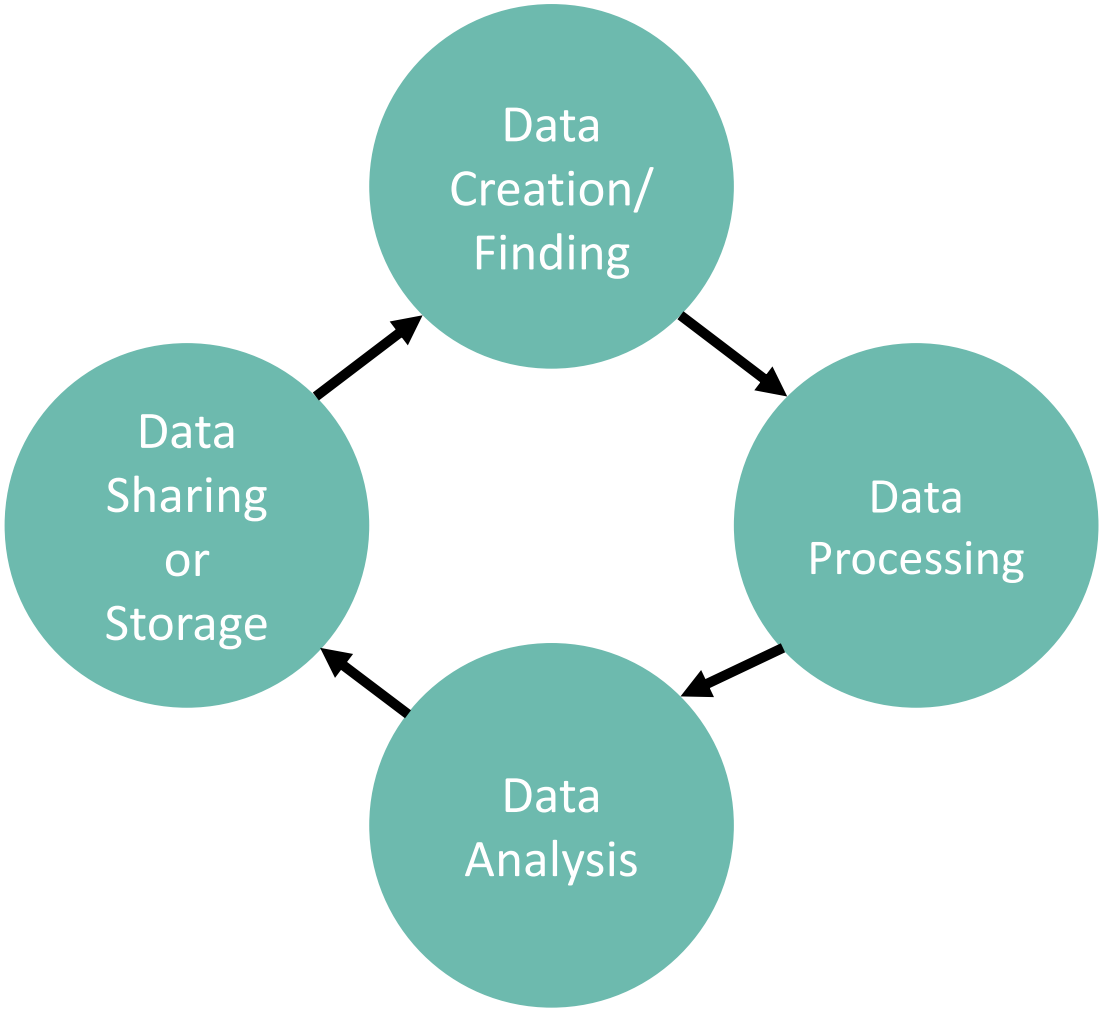
- Data use agreements (DUAs) and licenses
- Data security
- Data sharing requirements

Compliance

# Core concept: data literacy

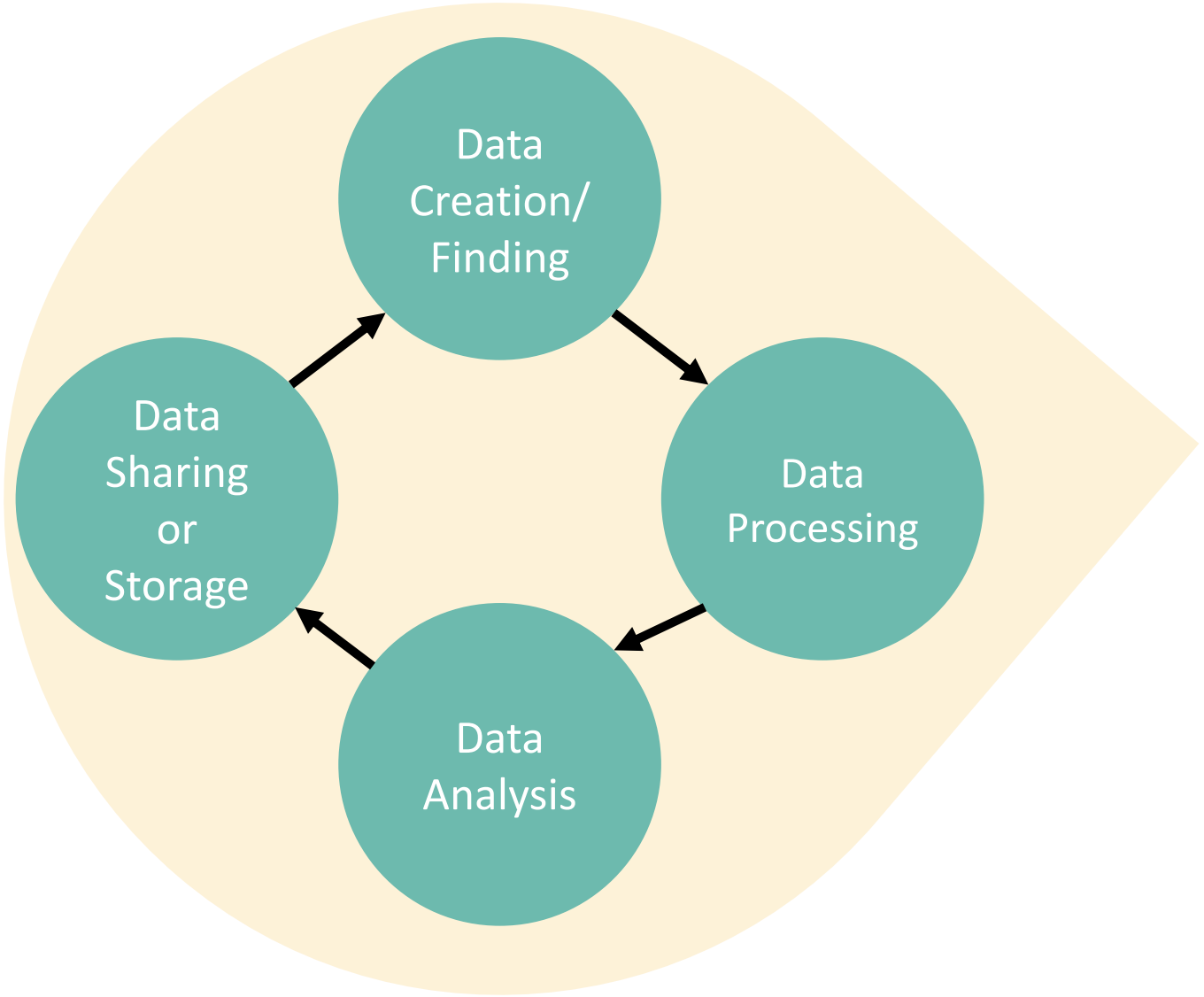
Yale UNIVERSITY LIBRARY
*Harvey Cushing/John Hay Whitney Medical Library*

# Research Data Lifecycle




The research data lifecycle model maps out a research project with a data-focused perspective.

# Research Data Lifecycle




- Data Security
- Documentation
- Operational Data Storage

# Data Sourcing

- Who collected the data?
- How was the data collected
- Why was the data collected?
- Who is providing the data?
- Who else has used this data for research?
- What is included in this dataset (and what is omitted)?

# Finding or collecting data

What do you need to know to fully understand the context of your data

Yale UNIVERSITY LIBRARY
*Harvey Cushing/John Hay Whitney Medical Library*

# Collecting Data

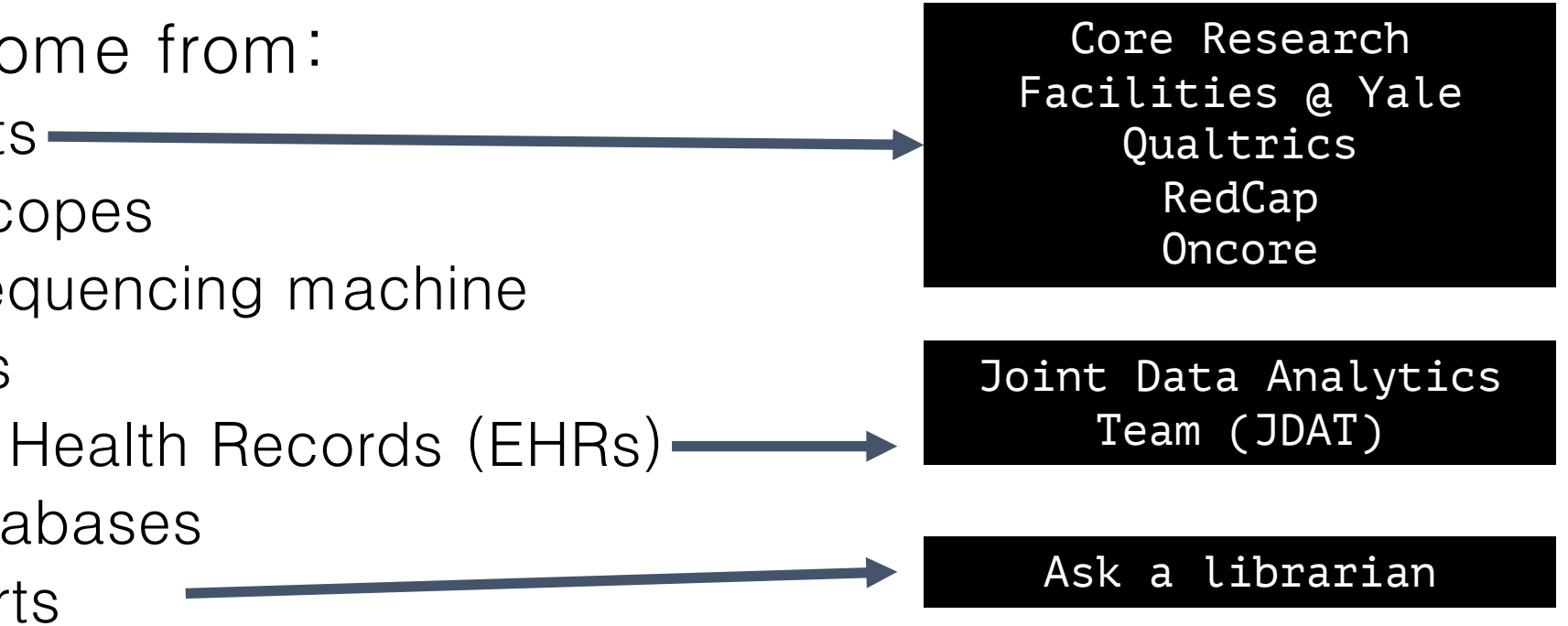
Data might come from:

- Instruments
  - Microscopes
  - RNA sequencing machine
  - Surveys
- Electronic Health Records (EHRs)
- Online databases
- Web exports

# Collecting Data

Data might come from:

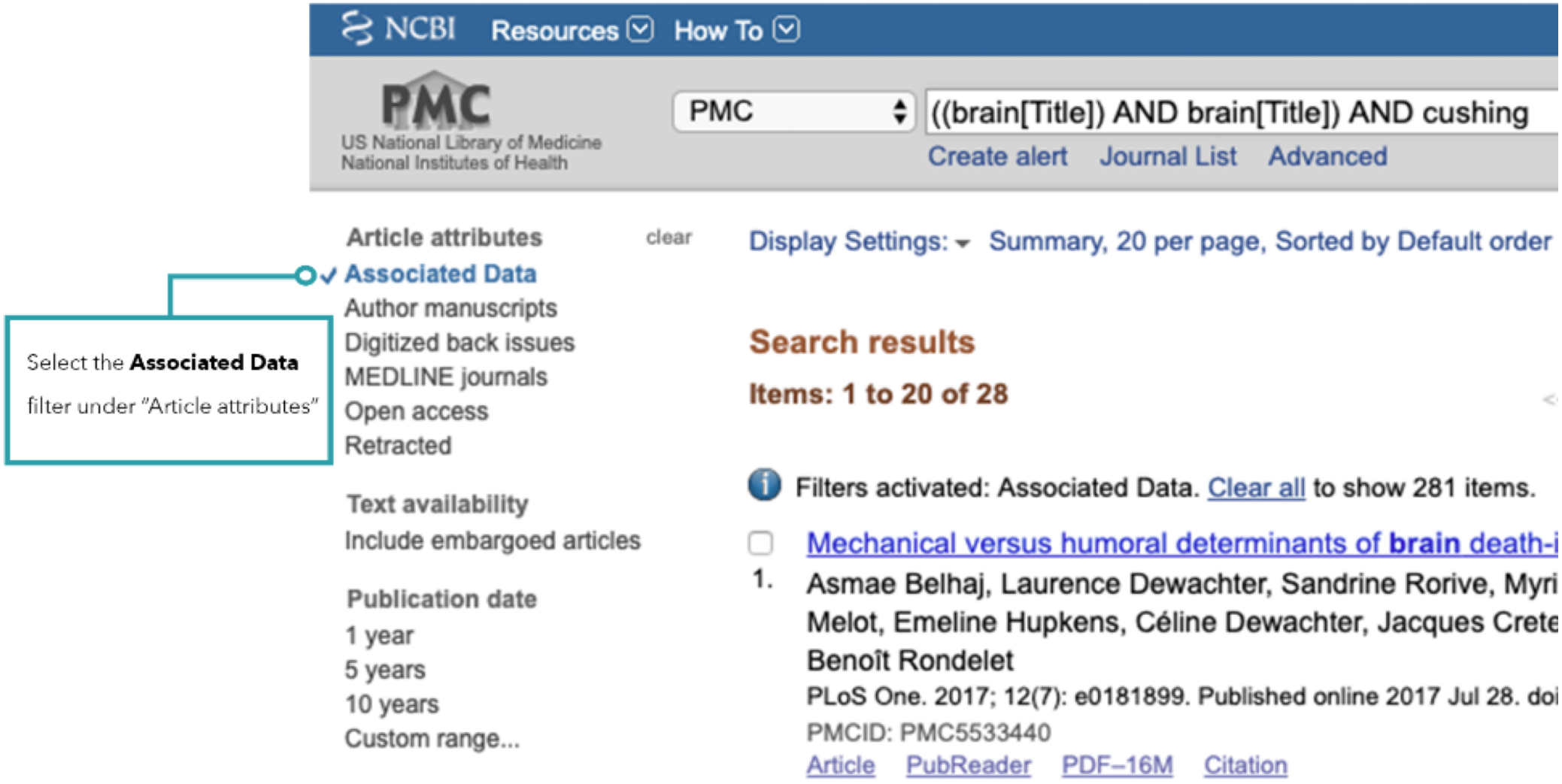
- Instruments
  - Microscopes
  - RNA sequencing machine
  - Surveys
- Electronic Health Records (EHRs)
- Online databases
- Web exports



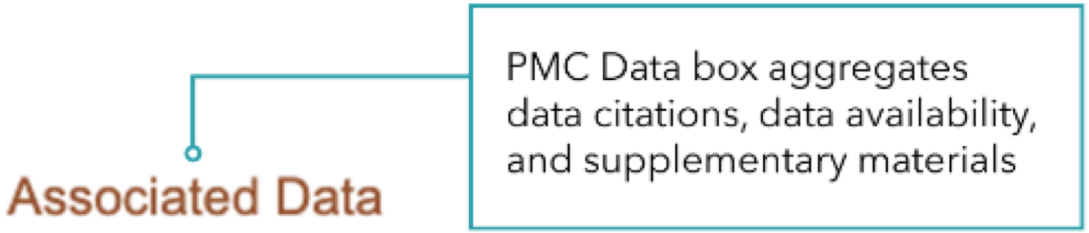
# Finding Data

- From literature
  - Supplemental material
  - Data availability statements
  - Data citations to databases or datasets
- From data repositories
- From research organizations

# Filtering for data in PubMed Central



# PubMed Central ”Data Box”




▾ Supplementary Materials

**S1 Data:** Datasets used and analyzed in the current study. Brain death (BI (CO), systemic arterial pressure (SAP), right atrial pressure (RAP), Noradr pressure (PAWP), pulmonary artery pressure (PAP), pulmonary vascular re 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 adhesic 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.

# Data Repositories

- Data repositories store and share datasets
- Search for repositories at re3data.org
- Find discipline specific data repositories Use the Cushing/Whitney Medical Library Data Quick Search Tool

# Library Data Quick Search

## Data Source Quick Search

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

**Search**

SEARCH RESET

| DATA SOURCE | DESCRIPTION | TOPICS |
| --- | --- | --- |
| CDC Data Catalog | Datasets and data visualizations from the Centers for Disease Control and Prevention. | • Administrative data<br>• Biomonitoring<br>• Disability & health & toxicology<br>• Injury<br>• Vaccination<br>• Violence<br>• Pregnancy<br>• Disability & health |

# Organizing Data

# Naming files

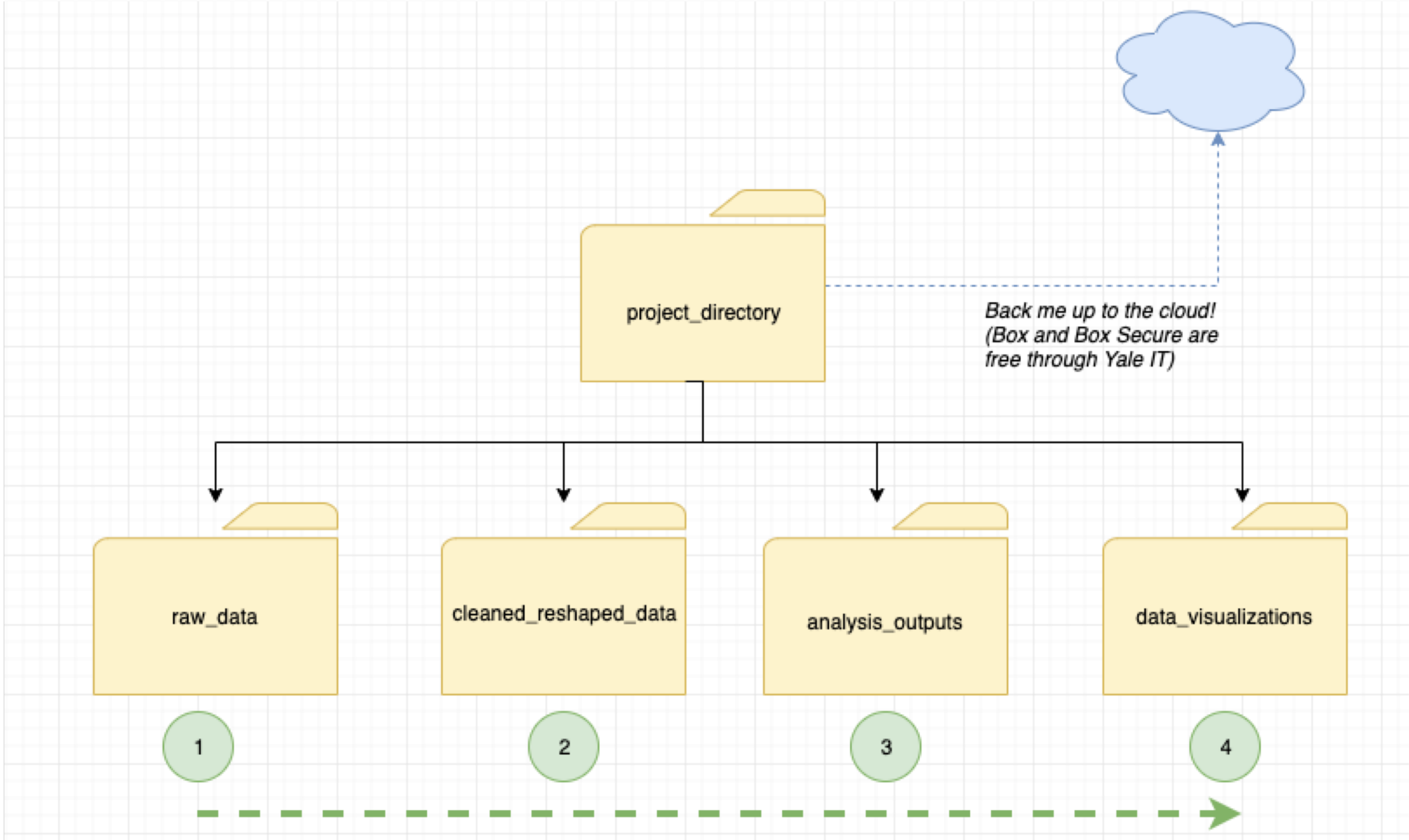
- File names should be consistent
- Short but descriptive (<25 characters)
- Use date format ISO 8601: YYMMDD
- Include a version number
- Document your file naming rules

# Spreadsheet Organization

- Columns contain variables (the measured elements)
- Rows contain observations
- Each cell should contain only one element
- Keep “master copies” of your raw data (create new Excel tabs or files as necessary)

# Folder organization

- Organize your folders in a way that makes sense to you
- Folder structures should be wider than they are deep
- Consider using folder organization template
- Documentation can capture organization rules


project_directory
Back me up to the cloud!
(Box and Box Secure are
free through Yale IT)
raw_data
cleaned_reshaped_data
analysis_outputs
data_visualizations
1
2
3
4

# Documentation

# Documentation – what is it?

Documentation is 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

# Types of research documentation

## ReadMe Files

Describe local files and folders

## Data Dictionaries

Describe data in spreadsheets or databases

## Codebooks

Describe survey design, variable meanings, and coded data

# Wrangling Data

# Stages of Data Wrangling

1. Selection: match data to 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 |

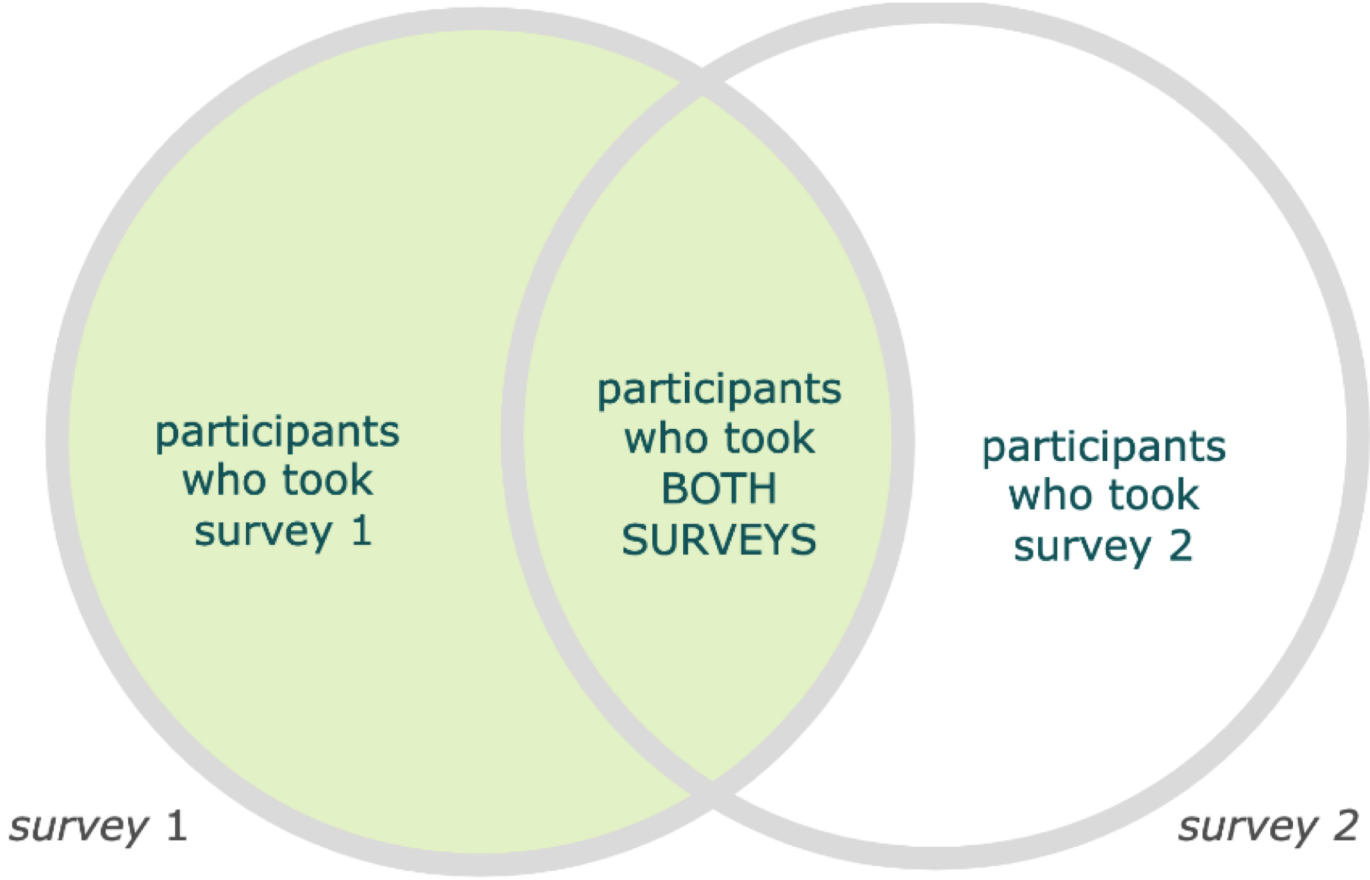
| last-name | first-name | question-1-a | question-1-b | question-1-c | question-1-d |
| --- | --- | --- | --- | --- | --- |
| Smith | Jane | a | b | NA | d |
| Doe | John | NA | b | c | NA |

# 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



# Core concept: technological considerations

# Data Sharing

## Transferring working data

- Yale Secure File Transfer (a.k.a. Filelocker)
- Yale Box and Secure Box
- Yale encrypted email services

box

Yale Secure File Transfer

## Sharing data with a wider audience

- NIH data repositories
- Disciplinary data repositories
- Dryad @ Yale

DRYAD

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

# Working with Data Files

- How large are your files?
- What environment do you need to use to use the file?
- Is your computer able to open the file?
- Will you need to use high performance computing?
- What software will you use for analysis?
- What skills do you need in order to work with your data?

# Data Security

- Protect data against loss due to accidental or technical problems
- Encryption, use of secure software, use and protect passwords
- Use backup services and save files in multiple places
- What are ITs data security services?
  - cybersecurity.yale.edu/protectyourdata

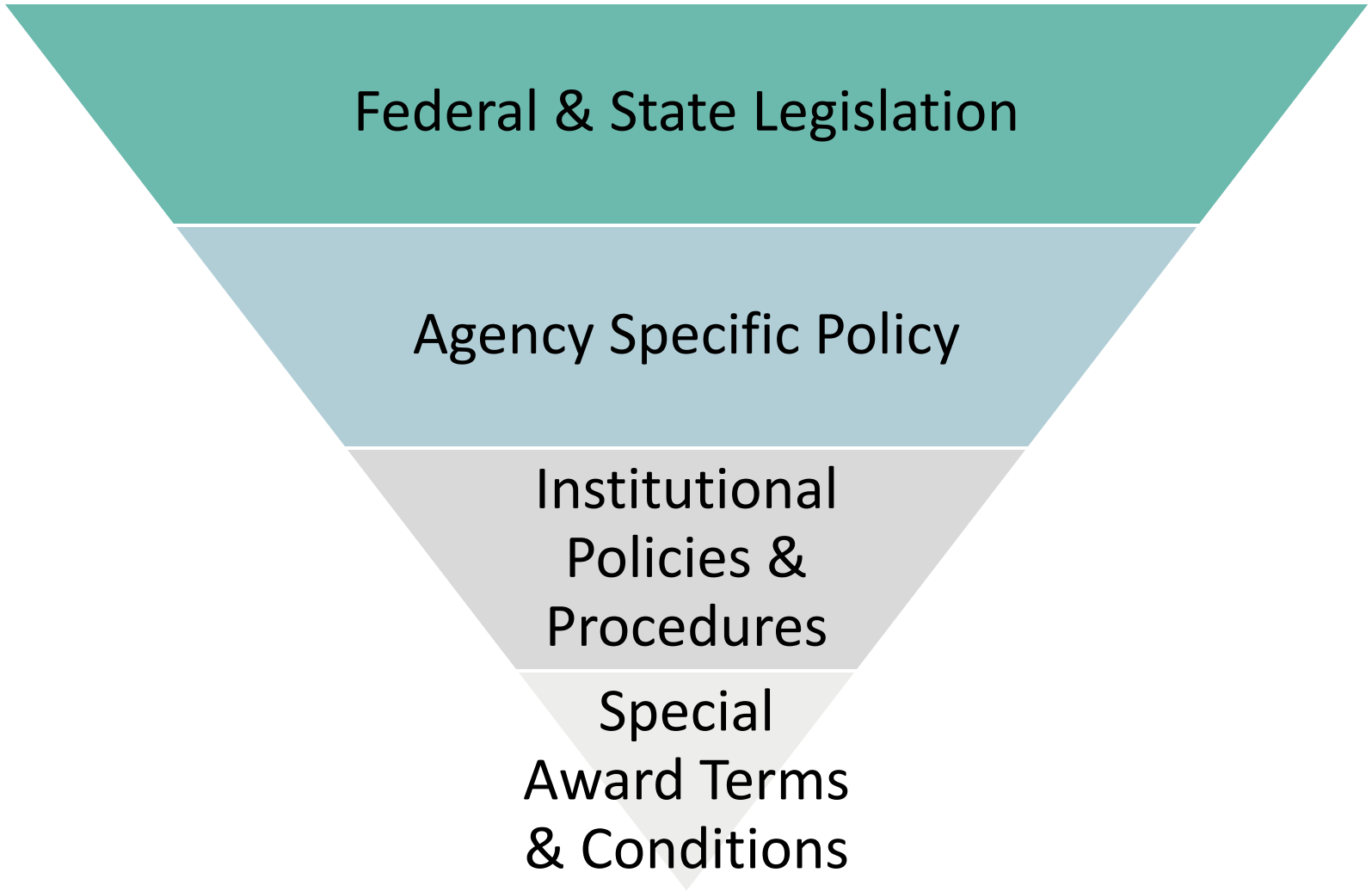
# Data Storage

- If your data low, moderate or high risk?
- How actively are you using your data?
- Is this your data or your lab's data?
- Is there a data usage agreement that states how data should be handled?

Data Storage Options @ Yale: docs.ycrc.yale.edu/data

# Core concept: compliance

# Levels of Policy Compliance



# Information on Data Policies

- RDM Policies: https://guides.library.yale.edu/rdm_healthsci/policies

- Yale Research Data & Materials Policy: https://your.yale.edu/policies-procedures/policies/6001-research-data-materials-policy

# Common Policy Coverage Areas

- Data management policies
- Data sharing policies
- Data security minimum standards
- Data Use Agreements
- Data Licenses
- Human subjects data handling policies

# Research Data Services

@ the Cushing/Whitney Medical Library

Yale UNIVERSITY LIBRARY
*Harvey Cushing/John Hay Whitney Medical Library*

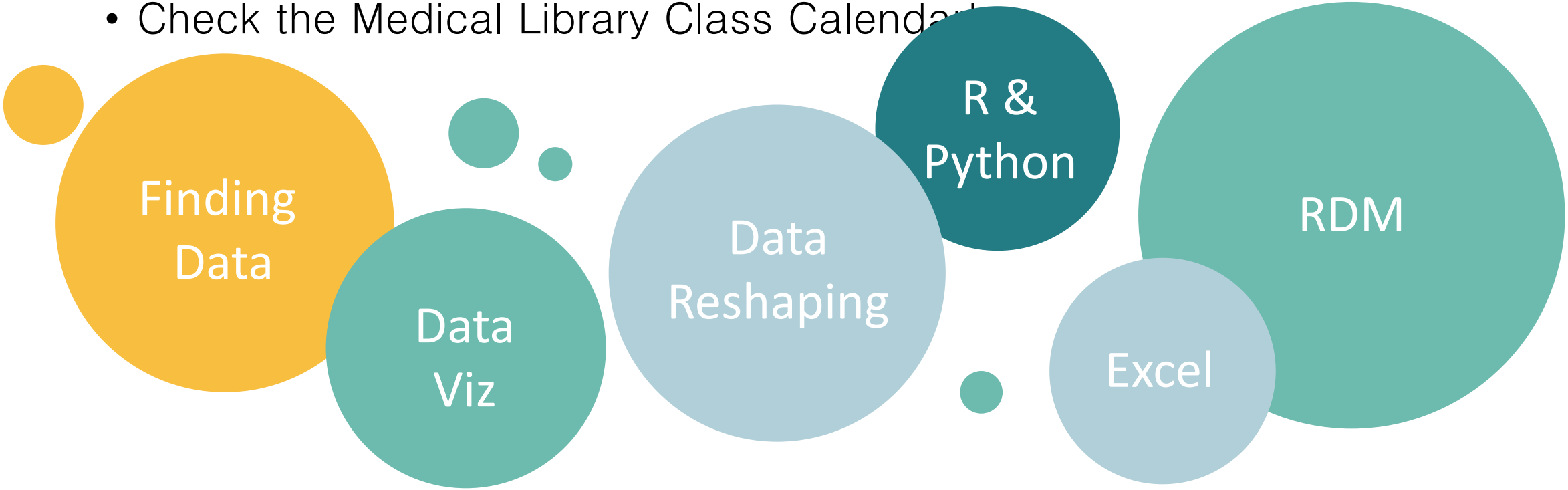
# Classes @ the Cushing/Whitney Medical Library

- Research Data Documentation
- Introduction to R
- Working with Data
- Unix Shell: Working with Data

https://library.medicine.yale.edu/research-data/classes-materials

# Office Hours

- Drop-in and ask questions about research data, research data management and data visualization
- Schedule changes on a monthly basis
  - Check the Medical Library Class Calendar



# Find More Research Help

- Bioinformatics Services (Medical Library)
- Statistical Support Services (CSSSI)
- Yale Center for Research Computing (YCRC)
- Research at Yale (research.yale.edu)
- Yale Center for Analytical Sciences (YCAS)
- Yale Center for Clinical Investigation (YCCI)


Yale UNIVERSITY LIBRARY

*Harvey Cushing/John Hay Whitney Medical Library*