SCHARE

Data Preparation for Creating an AI-ready Quality Data

May 21, 2025

Deborah Duran, PhD · NIMHD
Mark Aronson, PhD · NIMHD
Mohammad Arifur Rahman, MS · NIMHD

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	☐	☐	☐	☐
R	☐	☐	☐	☐
Cloud computing	☐	☐	☐	☐
Terra	☐	☐	☐	☐
Health disparities research	☐	☐	☐	☐
Health outcomes research	☐	☐	☐	☐

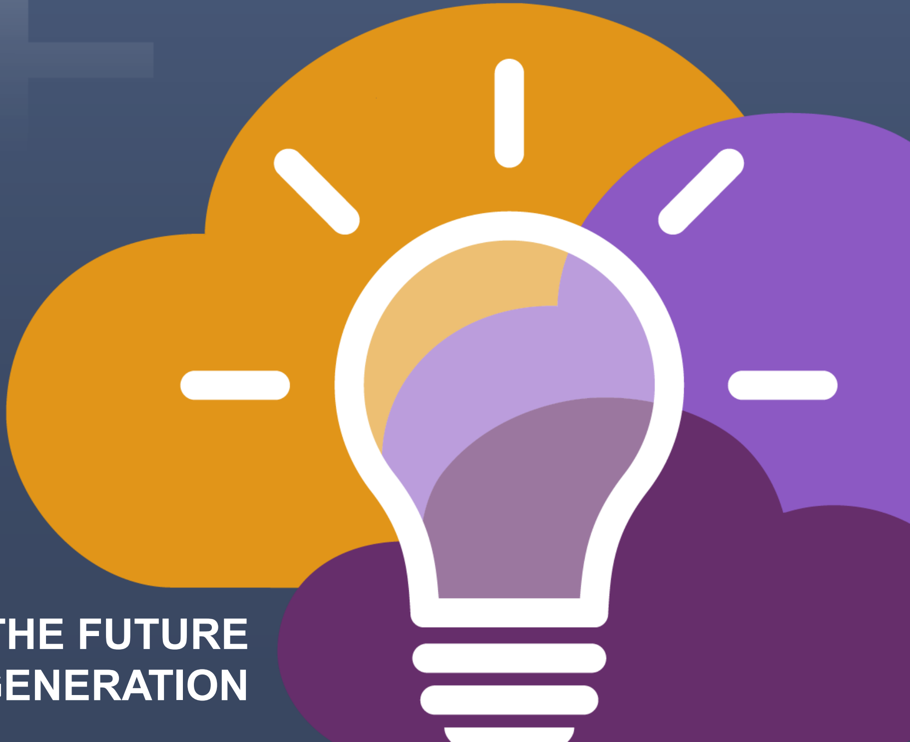
Interest poll

I am interested in (check all that apply):

- ☐ Learning about Health Disparities and Health Outcomes research to apply my data science skills
- ☐ Conducting my own research using AI/cloud computing and publishing papers
- ☐ Connecting with new collaborators to conduct research using AI/cloud computing and publish papers
- ☐ Learning to use AI tools and cloud computing to gain new skills for research using Big Data
- ☐ Learning cloud computing resources to implement my own cloud
- ☐ Developing ethical AI strategies
- ☐ Other

SCHARE

What is SCHARE?



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

SCHARE

Science
Collaborative for
Health disparities and
Artificial intelligence
Reduction of
Errors

Register: nimhd.nih.gov/schare

SCHARE is a cloud-based population science data platform designed to accelerate population health research, including chronic diseases, health disparities & health outcomes by utilizing transparent artificial intelligence (AI) approaches with a focus on the reduction of errors in the use and reuse of models designed to accelerate innovative research that includes place-based factors and biologics for whole-person health discoveries.

SCHARE aims to fill five critical gaps:

- Leverage population science, place-based, and behavioral Big Data and cloud computing tools to foster a paradigm shift in population health research to generate innovative whole-person health discoveries using AI
- Advance use of transparency and sophisticated inquiry to develop innovative strategies and differing perspectives to reproducibility and to reduce AI errors
- Upskill novice untrained users in data science through cloud computing skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Provide a data science cloud computing resource and data center for community colleges, and low resource institutions and organizations
- Offer a project data repository centered on core common data elements for enhanced data interoperability and compliance with NIH Data Management and Sharing Policy

SCHARE

Google Platform
Terra Interface

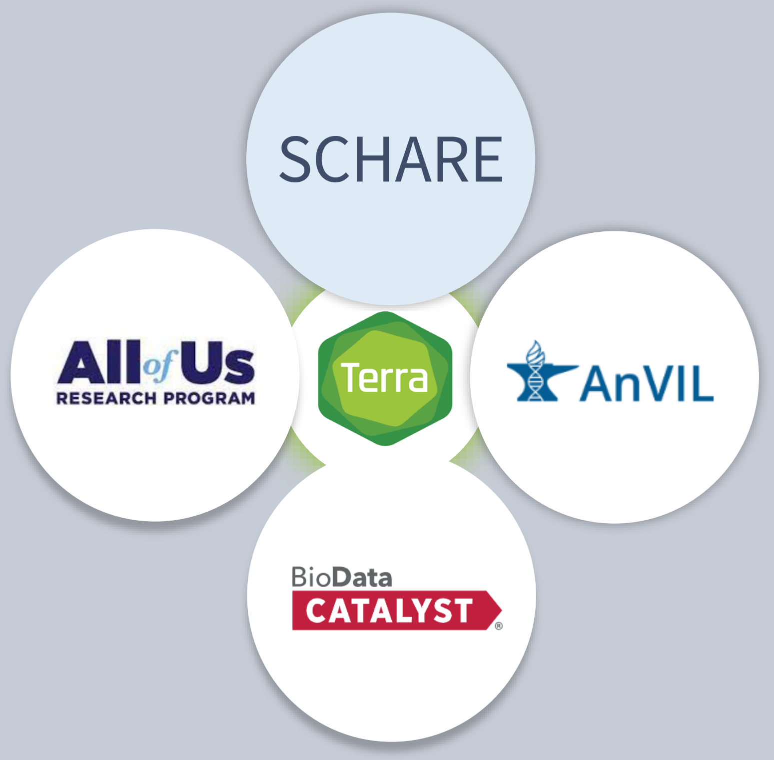
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on SCHARE prepares you to use other NIH platforms

Terra recommends using **Chrome**

Must have a **Gmail** friendly account

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms with Terra interface for collaborative research



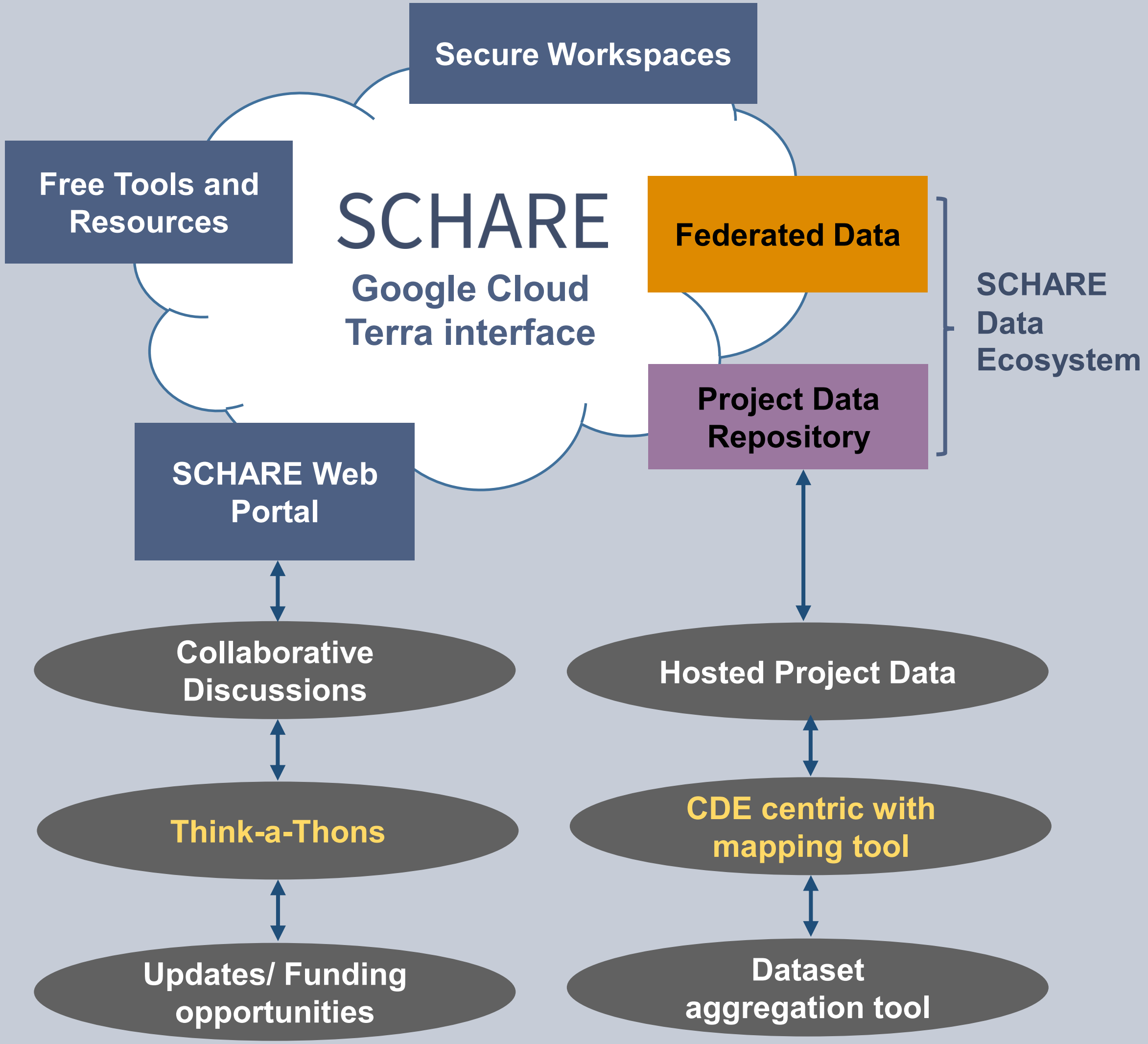
BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

SCHARE Components

SCHARE co-localizes within the cloud:

1. **Datasets** relevant to health disparities, health care delivery, and health outcomes research, including social determinants of health and other social science behavioral data
2. **A project data repository** for NIH-funded projects centered on Core Common Data Elements for enhanced data interoperability and compliance with NIH Data Management and Sharing policy
3. **Secure, collaborative workspaces** and for researchers and relevant collaborators
4. **Computational capabilities** for collaboratively evaluating designing and assessing fit-for-purpose utilization of datasets and algorithms to generate AI models that are effective and efficient

Intramural and Extramural Resource



SCHARE

SCHARE Terra Platform

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

National/Federated Datasets

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

SCHARE **Ecosystem structure**

Researchers can access, link, analyze, and export **a wealth of SDoH and population science related datasets** within and across platforms relevant to research about health disparities, health care delivery, and health outcomes, including:

300+
FEDERATED PUBLIC DATASETS

Public datasets

Publicly accessible, federated, de-identified datasets hosted by SCHARE or hosted by Google through the Google Cloud Public Dataset Program

Examples: *Behavioral Risk Factor Surveillance System (BRFSS)*
American Community Survey (ACS)

CDE
FOCUSED REPOSITORY

Project datasets

Publicly accessible and controlled-access, funded program/project datasets using Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

Examples: *Forthcoming datasets such as the Jackson Heart Study (JHS)*

Innovative Approach:
CDE Concept Codes
Uniform Resource Identifier (**URI**)

SCHARE **Ecosystem**

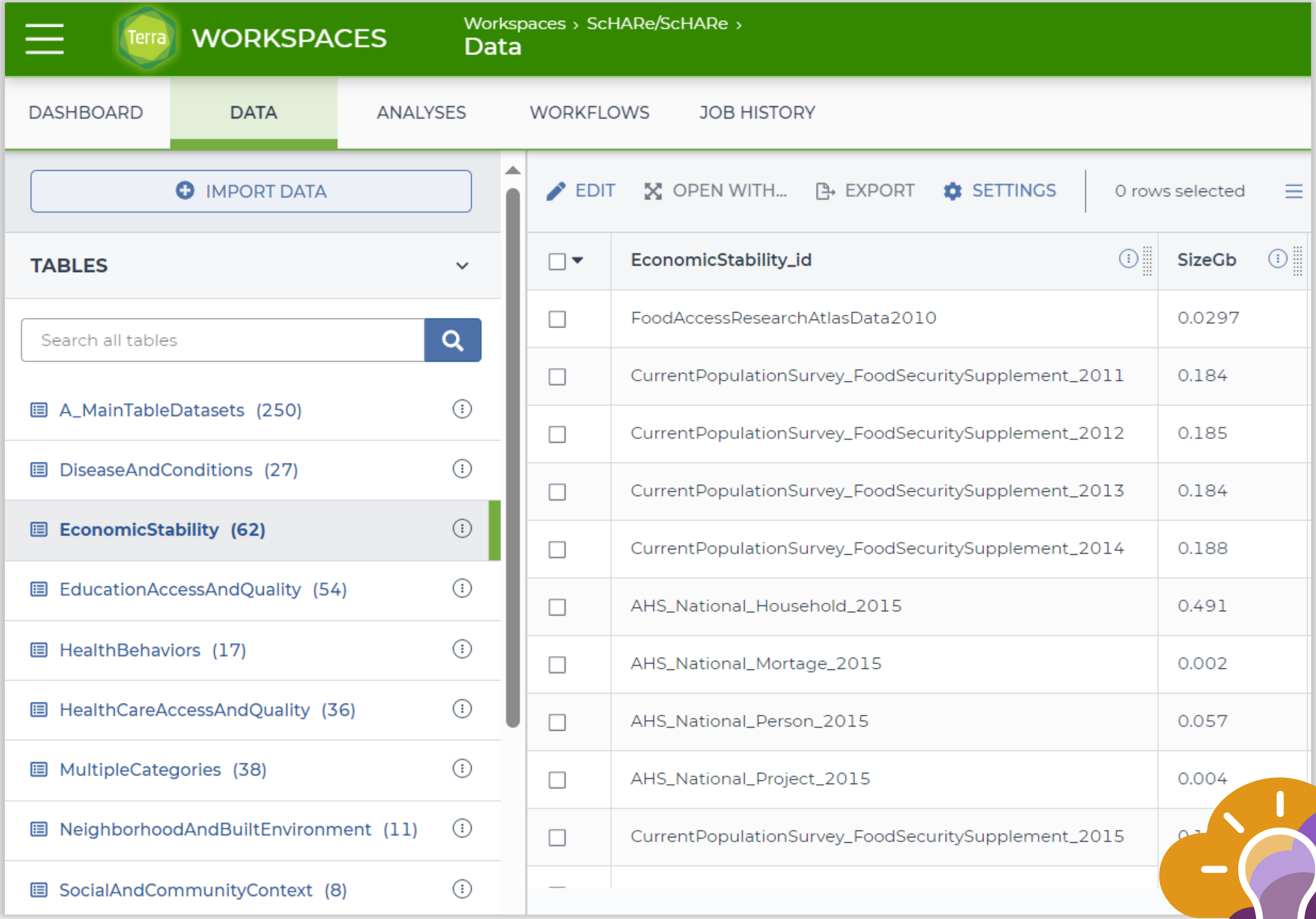
OVER 300 DATA SETS CENTRALIZED

Datasets are categorized by content based on the CDC **Social Determinants of Health categories**:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

with the addition of:

- **Health Behaviors**
- **Diseases and Conditions**



EconomicStability_id	SizeGb
FoodAccessResearchAtlasData2010	0.0297
CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.184
CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.185
CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.184
CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.188
AHS_National_Household_2015	0.491
AHS_National_Mortage_2015	0.002
AHS_National_Person_2015	0.057
AHS_National_Project_2015	0.004
CurrentPopulationSurvey_FoodSecuritySupplement_2015	[illegible]

SCHARE Ecosystem: Public datasets

Examples of interesting datasets include:

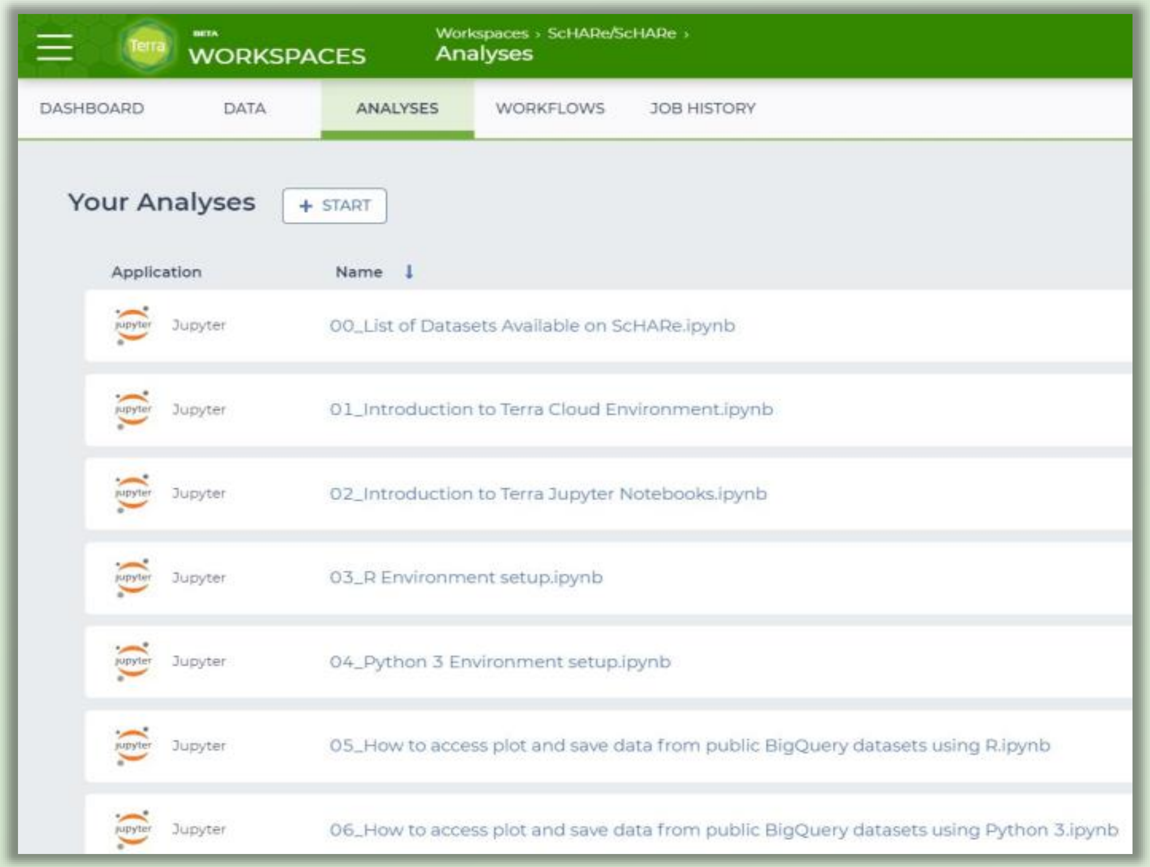
- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **U.S. Chronic Disease Indicators** (U.S. Census Bureau)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **National Mental Health** (SAMHSA)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **National Mental Health** (SAMHSA)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **BRFSS** Behavioral Risk Factors
- **Community Resilience Estimates:** Community resilience estimates calculated by modeling individual and household characteristics
- **Adult Indicators for Oral Health** (NOHSS)
- **Alzheimer’s Disease and Health Aging Data** (NIH)

Data Analytic and AI Tools

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

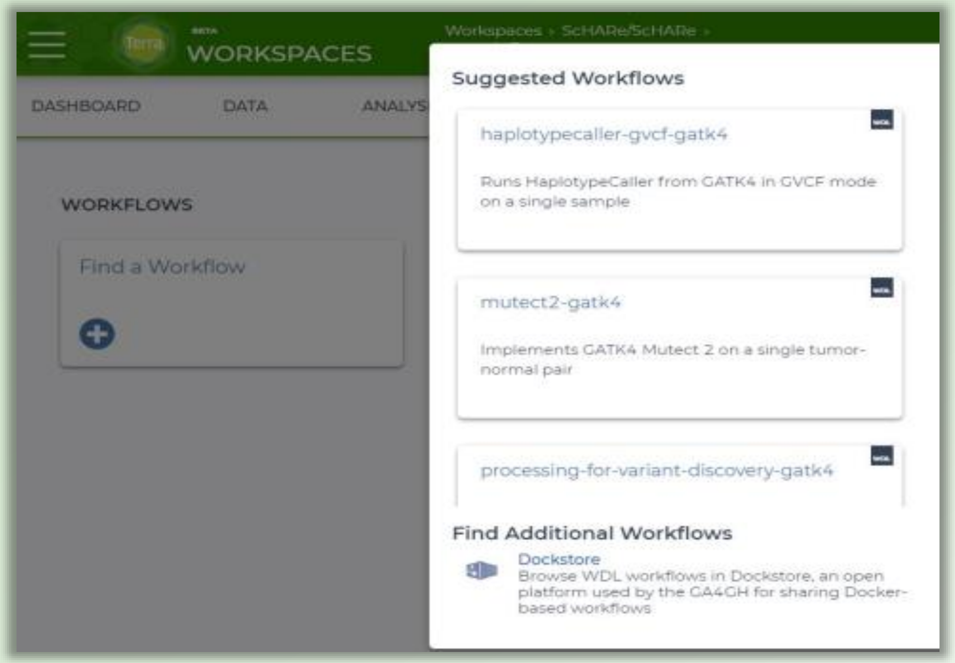
SCHARE Terra interface: Analyses (Notebooks)

Notebooks for analytics and tutorials



Modular codes

- **Easy-to-use copy and paste analytics**



- Modular codes developed for reuse

Data in SCHARE Repository Analyzed in SCHARE Terra

SCHARE Model Notebooks under Analysis Tab

Python

Copy & paste code for accessing datasets hosted on the SCHARE workspace

b. 00_Introduction to Python.ipynb

"Table of Contents"

Describes the purpose of all other notebooks in this section

b. 01_Python 3 Environment setup.ipynb

Describes what a python environment is and copy & paste code to set yours up

b. 02_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

Copy & paste code for accessing datasets hosted by Google BigQuery

b. 03_How to access plot and save data from ScHARe hosted datasets using Python 3.ipynb

Terra

Python code model notebooks (SCHARE Workspace -> Analyses -> Section B)

b. 04_How to upload access plot and save data stored locally using Python 3.ipynb

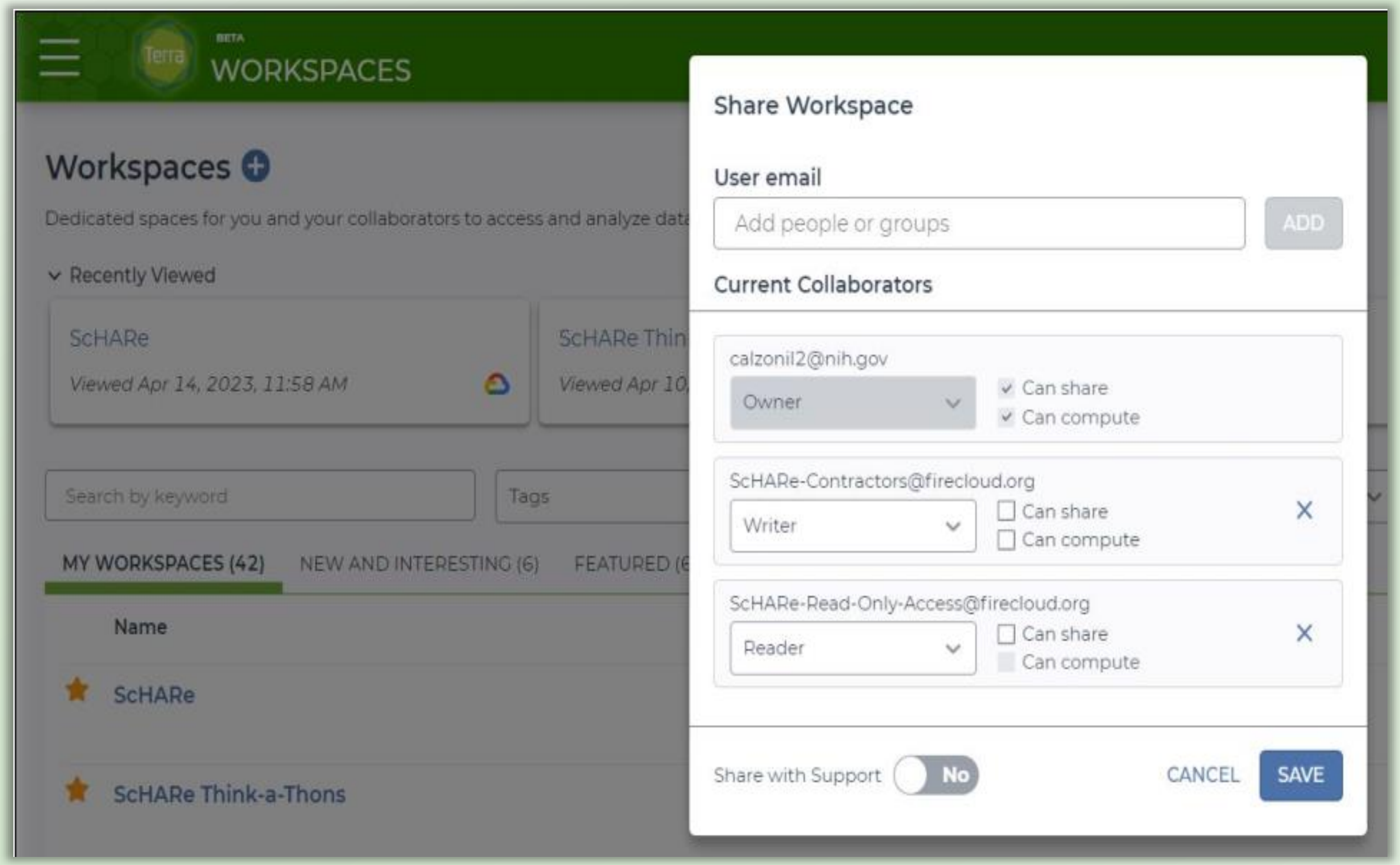
Copy & paste code for accessing data on your local computer

R

Secure Workspaces for Single and Collaborative Research

BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

SCHARE Terra interface: secure workspace



- Secure workspace **for self or collaborative research**
- **Assign roles**: owner, writer, reader
- **Host own data and code**
- **Own project costs**

SCHARE

Data Repository

NIH Data Management and Sharing Policy

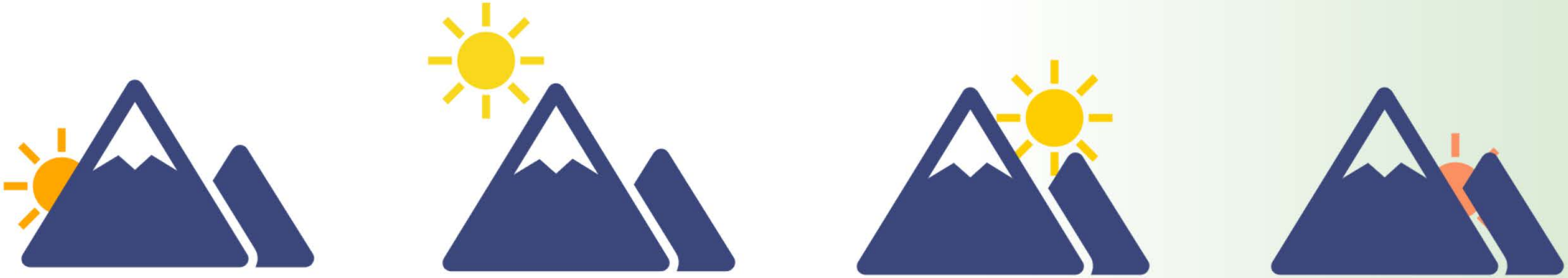
BE A PART OF THE FUTURE
OF
KNOWLEDGE GENERATION FOR POPULATION HEALTH

SCHARE Data Repository

PUBLICLY AVAILABLE SPRING 2025

The Four Data Lifecycle Stages

The SDR is here to support your research and your data throughout all stages of the data lifecycle. Our touchpoints can be contextualized by thinking about your data in these four stages.



Dawn

Dataset Creation
Researchers can choose to store their data themselves, uploading it upon study completion, or use the SDR as a storage interface.

Midday

Dataset Submission
Researchers submit their data for public sharing on the SDR, creating a controlled-access version if the dataset contains sensitive information.

Golden Hour

Dataset Access
Researchers use the public version of datasets on the SDR, or request access to controlled-access data, for secondary studies.

Sunset

Dataset Archival
Once the dataset meets the archival requirements, the dataset collection is removed from the SDR, and the underlying data is archived.

Key Features of the SCHARE Data Repository

Upload your own data

Store collected data and annotate with a data dictionary. Align data to the ScHARe CDEs.

Harmonize data to CDEs

Map uploaded data to CDEs. Join your data with project or federated data via CDEs.

Browse for data

Find relevant federated national datasets or other project data.

Manipulate and aggregate data for analysis

Filter, sort, and select subsets for specific purposes. Link and aggregate datasets.

Control privacy levels and data sharing

Share confidential data among colleagues. Share public access data with the research community.

Data Analysis via SCHARE Terra or local analysis platform

SCHARE CDEs Human & Machine Readable

Common Data Element (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection

Semantically Defined: (Human Readable)

Each are semantically defined by a standardized coding system for shared meaning

Use of international/national coding systems – LOINC, UMLS, SemNet, FHIR, NCIt

Alcohol: disinfecting or drinking?

Colon: sentence punctuation or biological organ?

Mole: animal, blemish, unit of measure, or spy?

&

Coded (Machine Readable) : Use URI approach of associated codes that can be mapped across coding systems to create data interoperability

Pipes to separate data points (i.e. flower| plant| succulent| grass| tree|)

Human & Machine Readable

SCHARE Core CDEs

PhenX Toolkit

NIH Endorsed

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex at Birth
- Marital Status
- Education
- Annual Household Income
- Household Size
- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self-Reported Health
- Health Conditions (and Associated Medications/Treatments)

- **NIMHD Framework***
- **Health Disparity Outcomes***

* Project Level CDEs

Available in Spanish

SCHARE has developed **Common Data Elements** to ensure consistent data collection across studies, facilitate interoperability, and link data from different sources

NIH CDE Repository: cde.nlm.nih.gov/home

PhenX Toolkit: www.nimhd.nih.gov/resources/phenx/

SCHARE SDR Collections & Associations

Collections

Each project establishes its COLLECTION:

- Own data (ongoing or final)
- Single or collaborative
- Data Documentation
- Privacy controls
- CDE mapping
- Metadata

Data Submission can be ongoing or at end of project.

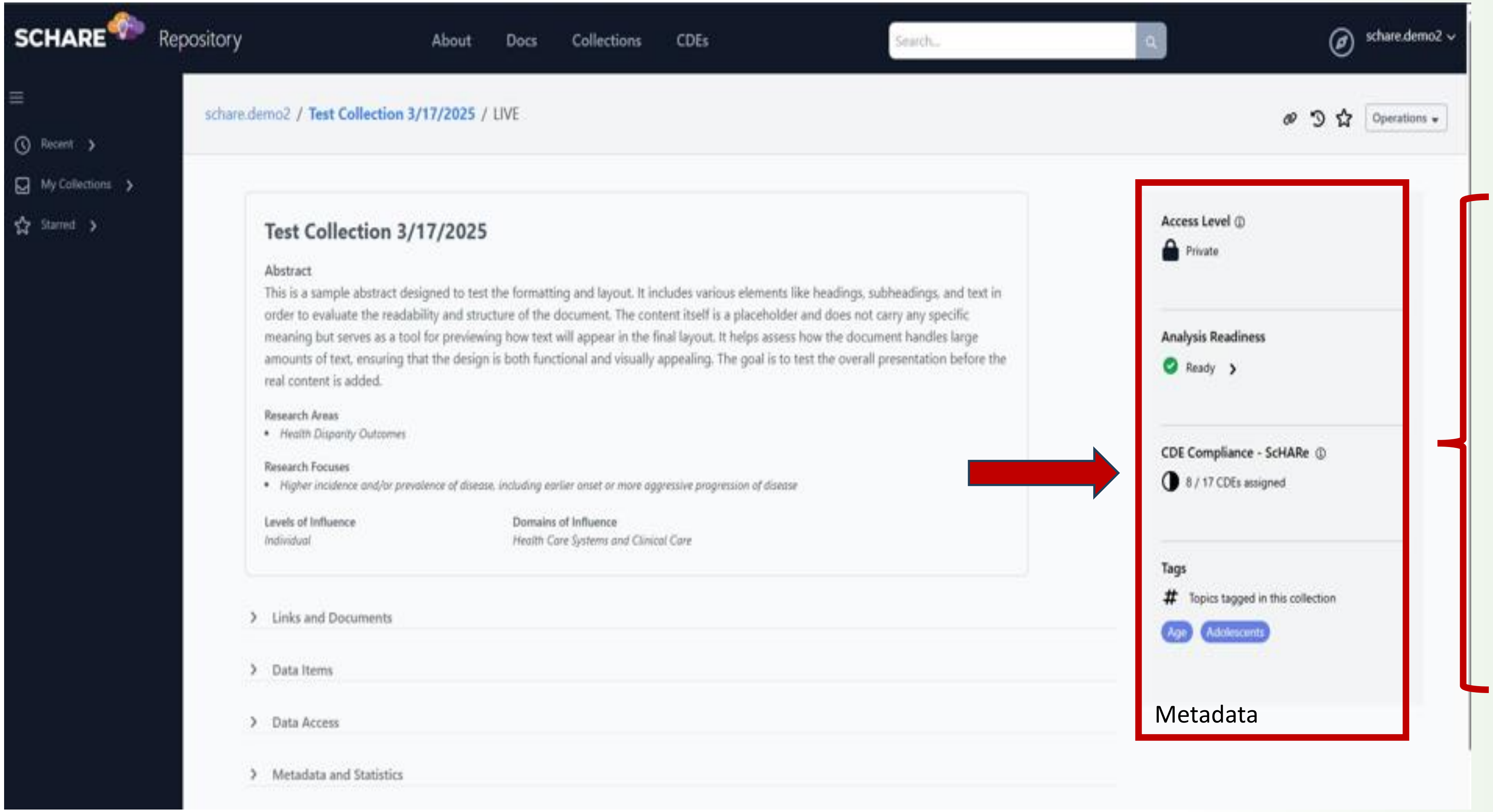
- **Can provide resource as a data center (ongoing)**
- **Fulfills Data Management and Sharing Policy (final)**

Associations

- ASSOCIATIONS comprised of multiple COLLECTIONs:
  - Creates parent collection
  - Own data (ongoing or final)
  - Single or collaborative
  - Data Documentation
  - Privacy controls
  - CDE mapping
  - Metadata
- Adds Collections to the Association

PUBLICLY AVAILABLE SPRING 2025

SCHARE Data Repository



Data Access & Data Readiness
Interoperability
Search
Metadata

By default, all collections start out as **Private**.

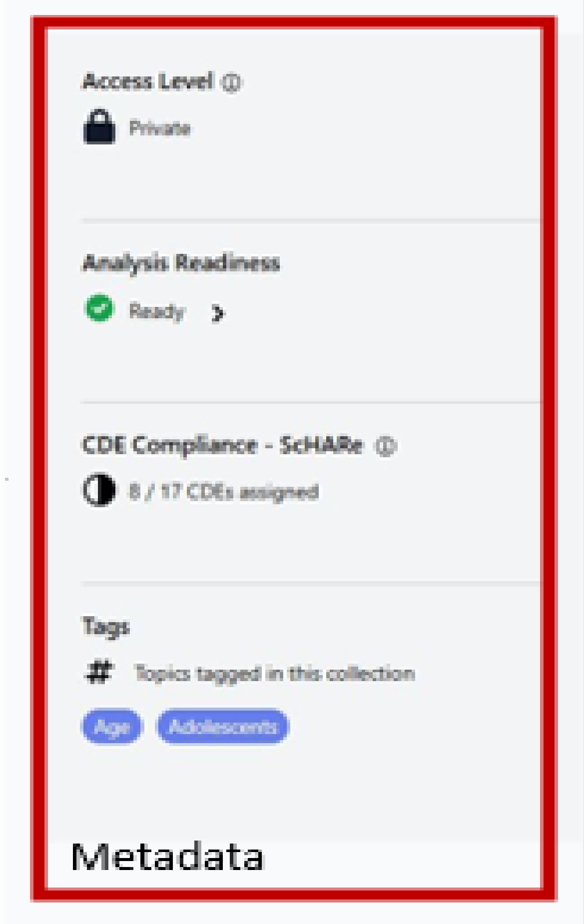
SCHARE Data Repository

Data Interoperability - CDEs

- All NIH funded data sets are mapped to the SCHARE CDE
- More mapped – better interoperability

By default, all collections start out as **None**

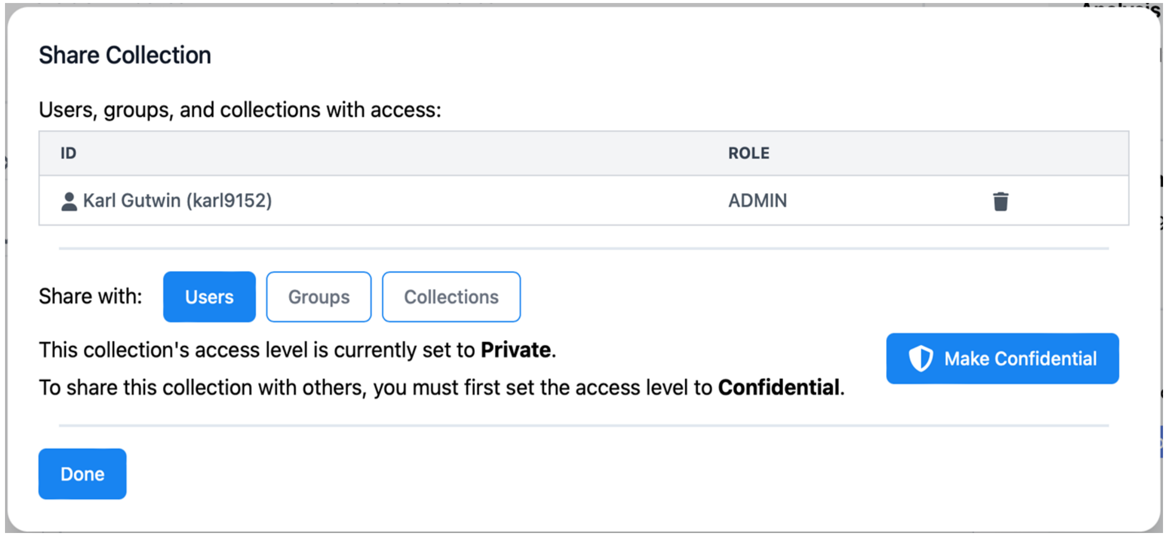
CDE mapping icon



SCHARE Data Repository

Access Levels and Sharing Data

The access level of a collection defines the maximum permissions that can be used to share it with others. You have control over how your data is shared on the ScHARe Data Repository.



By default, all collections start out as **Private**.

- **Private**: Only the collection's owner can access
- **Confidential**: The collection can be shared with named users
- **Controlled**: The collection can be shared with members of a controlled access group, as well as named users
- **Public**: The collection can be read by any user, including those not logged in; it can also be shared with named users

SCHARE **Data Repository**

Data Readiness

The readiness level delineates the preparation given a data set.

By default, all collections start out as **Raw**

- **Raw**: data not cleaned
- **Cleaned:** basic data cleaning, missing data, errors, labels, outliers etc
- **AI Ready:** includes all of cleaned data and imputations for missingness, aggregation of data sets for comprehensiveness, proxy variables, etc.

SCHARE AI Tools - Metadata

Metadata, and model documentation tools

```
import pyschare as sc
labels= sc.data_labels
```

Enter the project title and a brief description or abstract in the provided text boxes. Once done, press the 'Save' button to generate the dataset facts summary.

Project Title:

Project Description:

Save

Fill in the metadata fields with the relevant information about your dataset, such as filename, format, URL, and domain. After completing the fields, click the "Save" button to save and display the metadata table.

Filename:

Format:

URL:

Domain:

Keywords:

Type:

Geography:

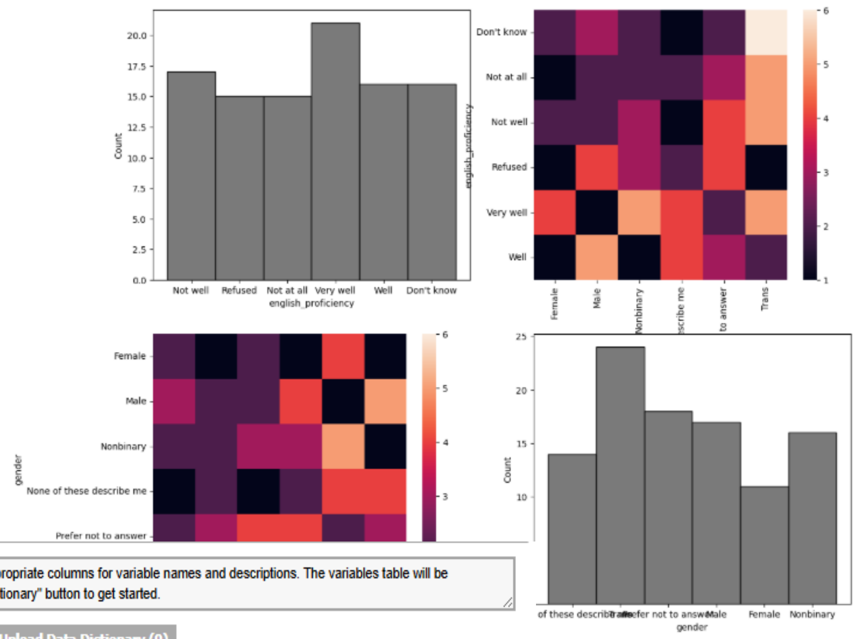
Data Collection Method:

Time Method:

Rows:

Choose two variables from the dropdown menus and click the 'Show Plots' button to create pair plots that visualize the distributions and relationships between the selected variables.

Variable A: english_proficiency Variable B: gender



Upload your data dictionary file (CSV format) and select the appropriate columns for variable names and descriptions. The variables table will be generated based on your selections. Click the "Upload Data Dictionary" button to get started.

Upload Data Dictionary (0)

Variable Names: Variable Descriptions:

Upload your dataset (CSV format), and click the 'Show Data' button to view the dataset. Using the dataset's variable names, input the names of ordinal, nominal, continuous, and discrete variables, separated by commas (e.g., Variable_a, Variable_b, Variable_c). After entering the variables, press the 'Show Statistics Table' button to generate and view the statistical summaries.

Upload Data (1)

Show Data

	Column1	entity:A_MainTableDatasets_id	Categories	Data	DataDictionary	FileFormat	Homepage	SizeGb	
0	0	2021FoodSecurityData	Health Care Access and Quality	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	XLSX	https://www.meps.ahrq.gov/mepsweb/data_stats/d...	0.806	Exp
1	1	2021FullYearConsolidatedData	Health Care Access and Quality	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	gs://fc-secure-d6e25d73-4b50-4dbc-ac10-ec68998...	XLSX	https://www.meps.ahrq.gov/mepsweb/data_stats/d...	0.118	Exp

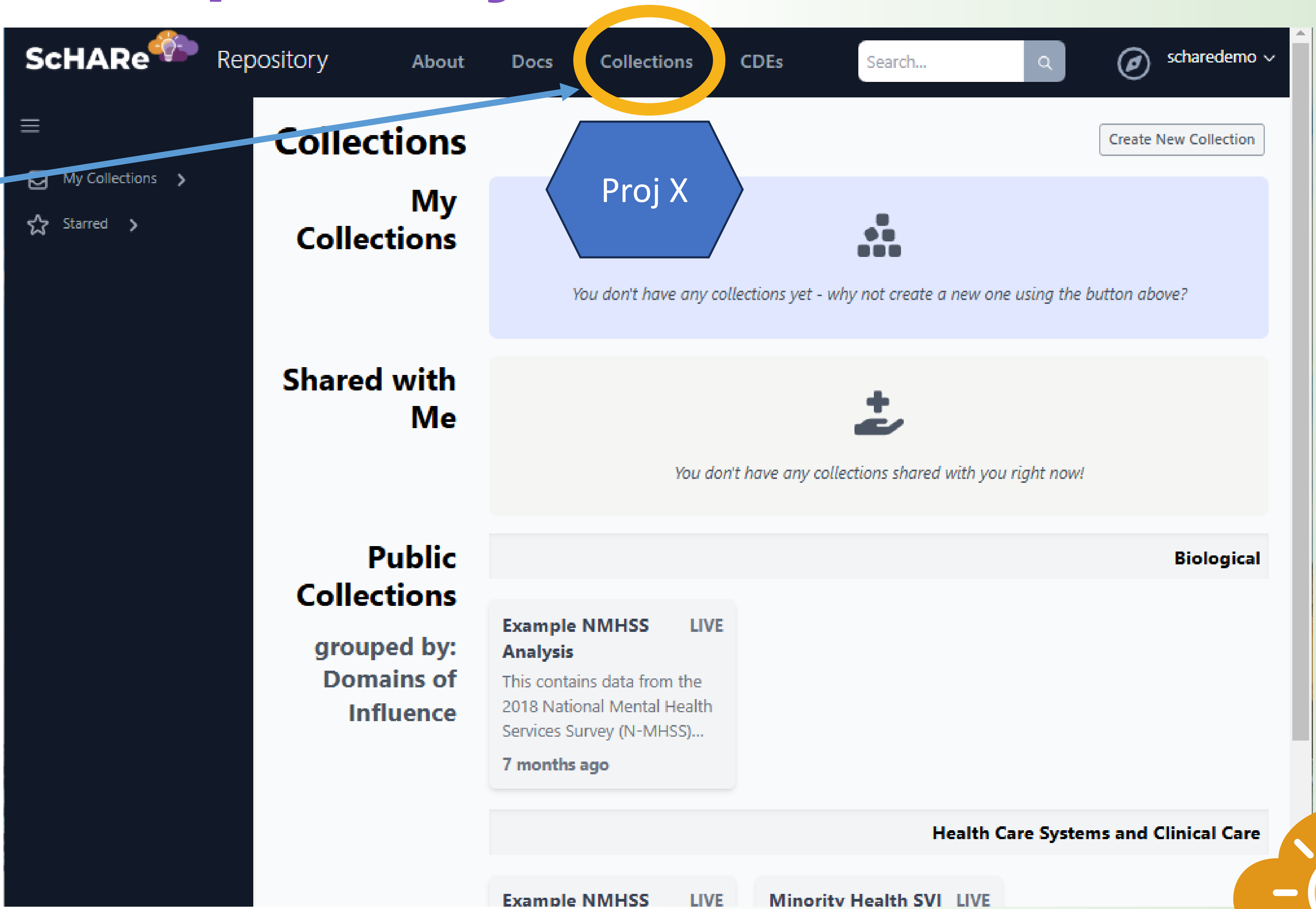
SCHARE Data Repository

Project Data

Collections are a place where you can describe and store your data and any related metadata and federated data.

Can be shared with colleagues

Privacy controls & published when you're ready.

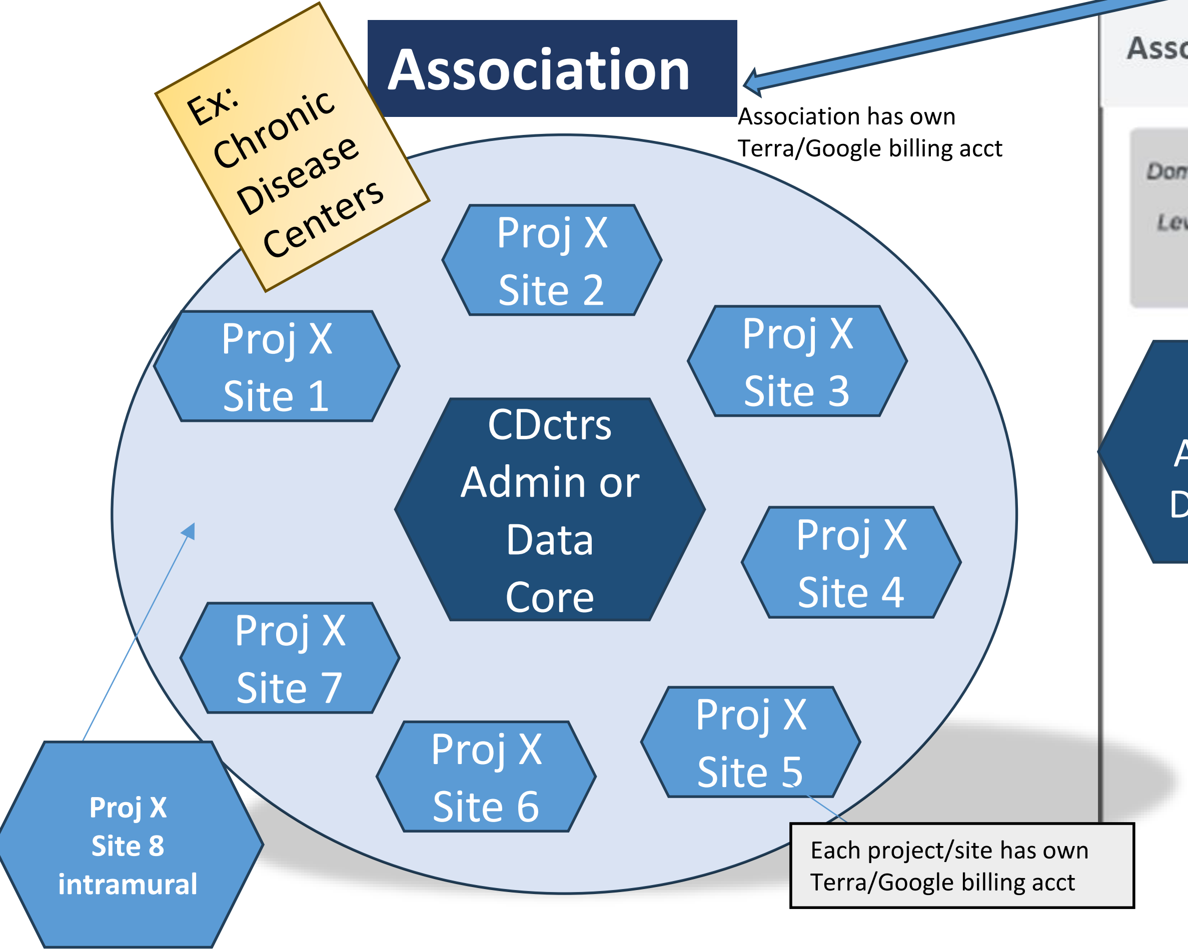


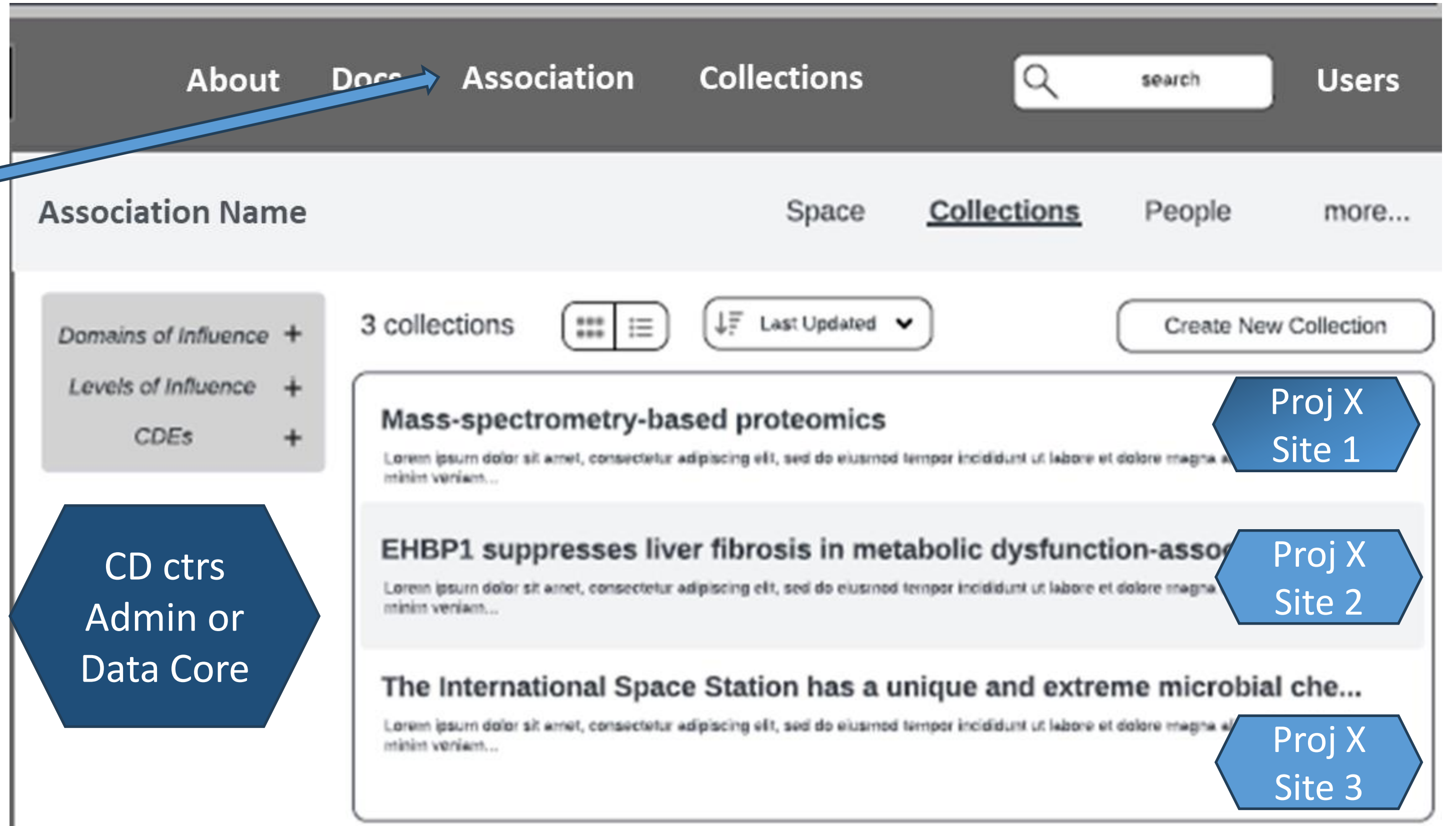
PUBLICLY AVAILABLE SPRING 2025

SCHARE Data Repository

Multi-Site Data

Project/Collections are integrated into the Association



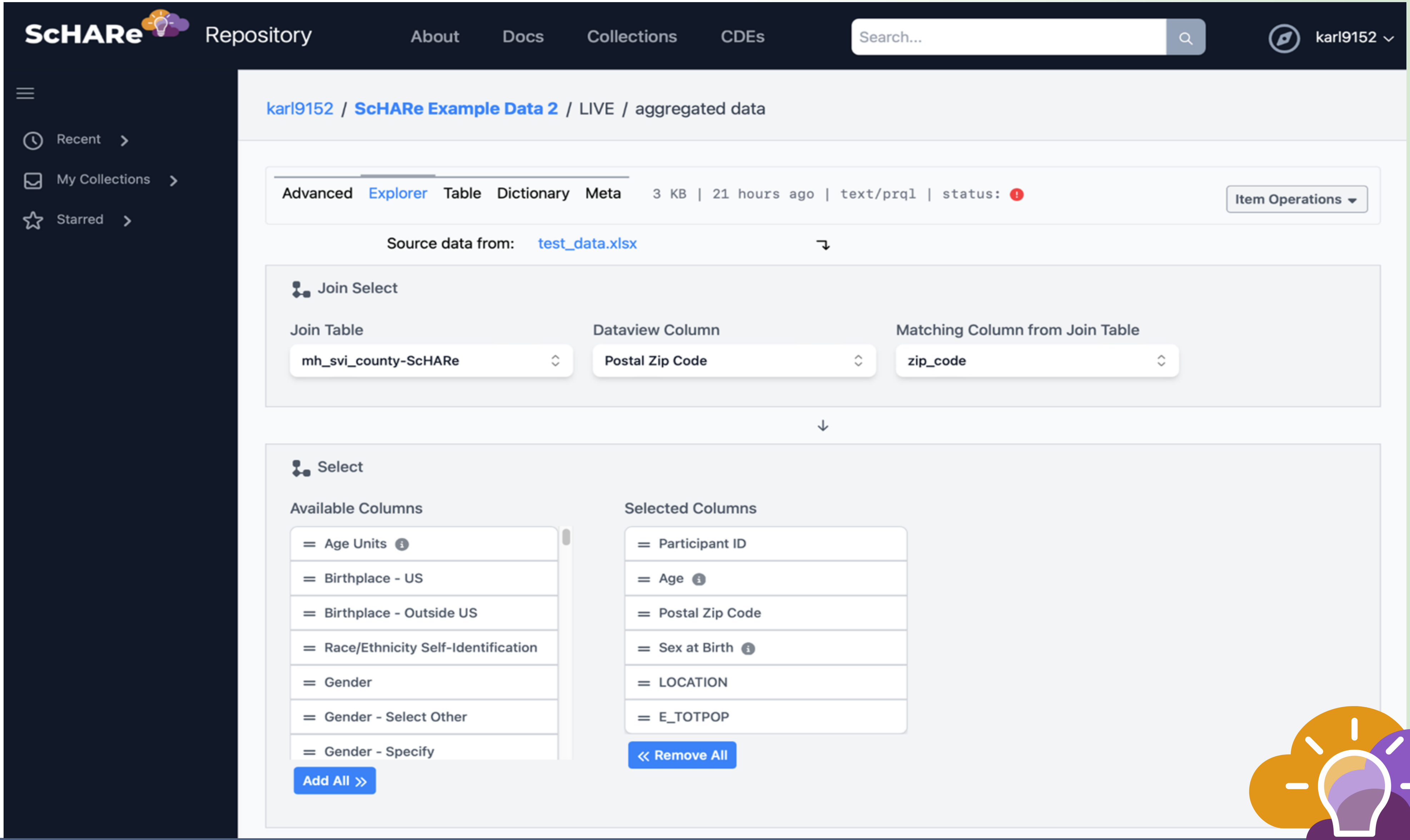


Each Association and Collection has its own space and metadata

SCHARE Data Repository

PUBLICLY AVAILABLE SPRING 2025

Data Aggregation Tool



SCHARE **AI Tools**

Authenticate

Gemini Assistant

Use **Gemini Assistant** to launch a simple Q&A chat window to get assistance with writing your data analysis code. The chat interface is powered by the Gemini model and is designed to answer questions related to assisting novice coders with writing analysis code. Type your question in the box and click the **Generate** button to call the model and generate an output.

Note: while the data you send through this tool and data sent back are protected under Terra's Enterprise Google Cloud permissions, and are not reused by Google for future model training, we advise not sending any sensitive information (e.g. PII or PHI) through the model. Sticking to general questions or inserting dummy variable names to your questions are good practices to ensure the privacy of your data.

Select Model Gemini 1.5 Flash

Select Location us-central1

Question: Type your coding question here...

Generate

SCHARE PySCHARE Python Package

PySCHARE package to search datasets and variables, subset, save, and visualize datasets

```
DataVisual()
```

Use the dropdown menus to select a dataset and configure your plot parameters.

- **Bar, count, box, boxen, strip, swarm, and violin plots** typically require a categorical variable on the X-axis (or hue) and a numeric variable on the Y-axis; see the categorical tutorial for details.
- **Scatter and line plots** call for numeric variables on both axes (e.g., time vs. measurement); refer to the relational tutorial.
- **Histograms** typically need a single numeric variable on the X-axis and are described in the distributions tutorial.

Use **"hue"** to differentiate categories by color, **"style"** to vary markers or lines, and **"size"** to scale markers based on another variable. The **"col"** and **"row"** options create subplots (facets) for comparison across categories, while the **"multiple"** parameter (e.g., "dodge," "stack," "fill") manages overlapping data displays.Once the plot type and settings are selected, click **"Show Plot"** to visualize the results.

Select Dataset

None
2021FoodSecurityData
2021FullYearConsolidatedData
2021JobsFileData
2021MedicalConditionsData
2021PersonRoundPlanPublicUseData
2022FoodSecurityData
2022FullCharacteristicsData
2022FullYearConsolidatedData

Select Plot

None
Bar Plot
Box Plot
Boxen Plot
Count Plot
Histogram
Line Plot
Point Plot
Scatter Plot
Strip Plot

Select X
Select Y
Select Hue
Select Style
Select Size
Select Column
Select Row
Select Layer — Layer

Show Plot

```
DataSubset()
```

Use the **Select Dataset** dropdown to choose a dataset. The available variables will be dynamically populated when you select options in the **Select Variables** dropdown. After selecting the desired variables from the **Select Variables** dropdown, you may visualize the data by clicking the **Show Data** button. This will display the first few rows of the specific columns selected in the output area below.

To save the displayed data, click the **Save Data** button. This action will store the selected data in your bucket and confirm the successful operation in the output area. Please make sure you have made selections in both the dataset and variables dropdowns before attempting to save.

Select Dataset

PLACES_500Cities_2021
PLACES_500Cities_2022
PLACES_500Cities_2023
PLACES_500Cities_2024
YRBSS_YouthRiskBehavior_2015
YRBSS_YouthRiskBehavior_2017
YRBSS_YouthRiskBehavior_2019
YRBSS_YouthRiskBehavior_2021
YRBSS_YouthRiskBehavior_2023

Select Variables

Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10

Show Data

Save Data

```
Calculate()
```

Use the **Select Dataset** dropdown to choose a dataset. The available variables will be dynamically populated when you select options in the **Select Variables** dropdown. After selecting the desired variables from the **Select Variables** dropdown, click the **Describe Data** button. This will display the summary statistics of the specific columns selected in the output area below.

Select Dataset

PLACES_500Cities_2021
PLACES_500Cities_2022
PLACES_500Cities_2023
PLACES_500Cities_2024
YRBSS_YouthRiskBehavior_2015
YRBSS_YouthRiskBehavior_2017
YRBSS_YouthRiskBehavior_2019
YRBSS_YouthRiskBehavior_2021
YRBSS_YouthRiskBehavior_2023

Select Variables

Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10

Describe Data

SCHARE

Research Think-a-Thons

Novice **training webinars** for data science, cloud computing and research using Big Data

Think-a-Thons

3rd Wednesday of every month 2 pm

Goals:

- Upskill novice untrained users in data science and cloud computing
- Foster a research paradigm shift to use Big Data in population health research, including health disparities/health outcomes
- Promote use of Dark Data (unused data epidemiologic studies)

1. TUTORIAL AND TARGETED THINK-A-THONS

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new/experienced users
- Networking
- Mentoring and coaching
- Topics include:
  - Data Science 101
  - Terra
  - Population: place- based factors
  - Common Data Elements
  - AI readiness
  - Transparent AI

Launched April 2024

2. RESEARCH THINK-A-THONS

- Multi-career (students to senior investigators)
- Multi-discipline (data scientists and researchers)
- Featured datasets with guest experts leads
- Guest experts in topic areas, analytics, data sources etc. to provide guidance
- Generate research idea - decide design, datasets and analytics
- Learn Ethical AI
- Publications

Register:

bit.ly/think-a-thons

SCHARE

Next Think-a-Thons:

bit.ly/think-a-thons

Register for SCHARE:

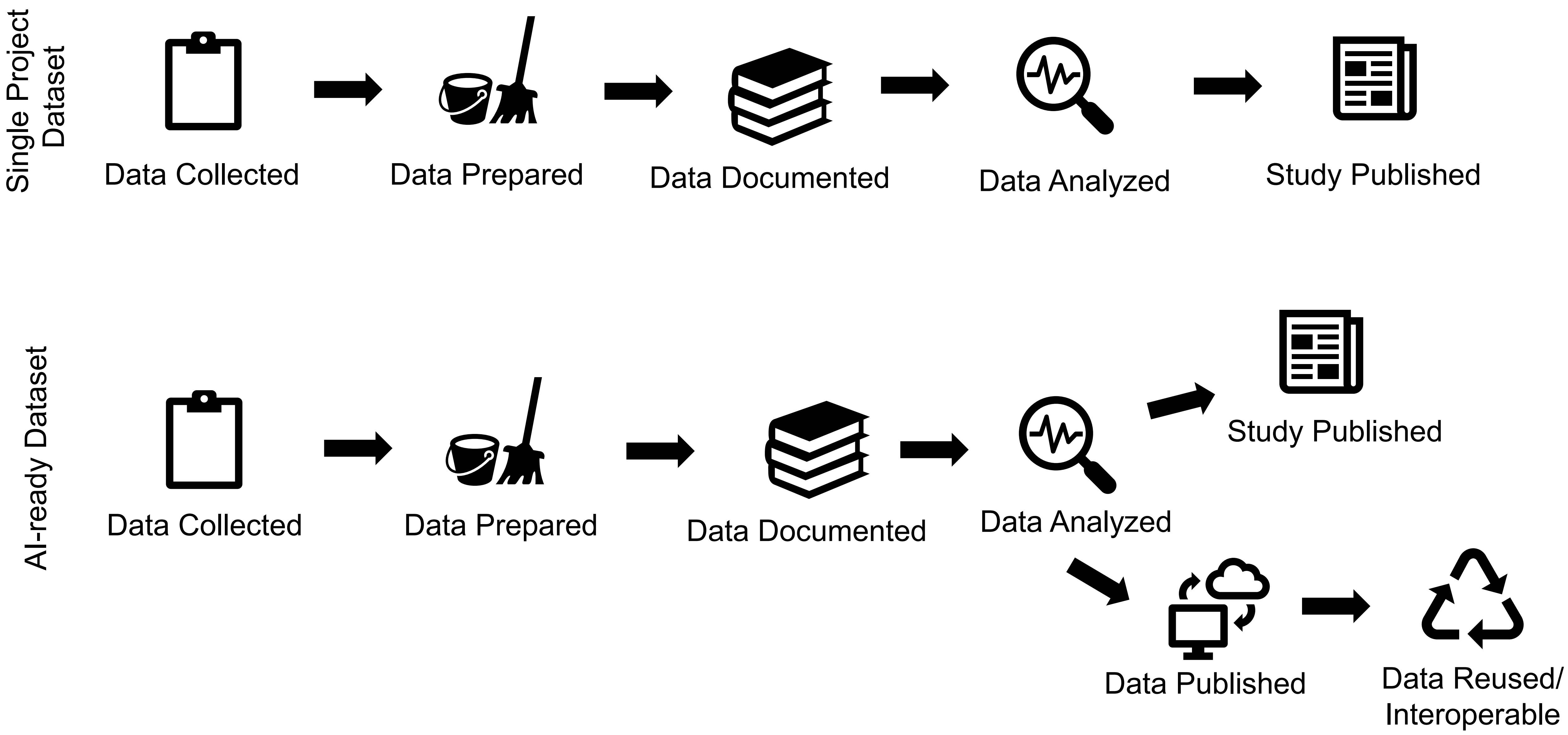
https://bit.ly/registerschare

schare@mail.nih.gov

The Importance of Data Preparation

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Preparation of AI-ready datasets enables data reuse



*We encourage all SCHARE members to create .csv files when uploading to the SCHARE Data Repository

Data preparation is critical for data-driven nature of AI models

AI models are compelling tools for research analysis because they…

“wide” data

 # features >> # observations

determine variable importance from many (100s-1000s) of variables

	Independence	Randomness	Normality	Linearity	Homoscedasticity	No multicollinearity
Statistical Models	✓	✓	✓	✓	✓	✓
Machine Learning Models	✓	✓	✗	✗	✗	✗

make fewer assumptions about the data going into them

But these advantages come with drawbacks because…

Observations (data) → Learn patterns between variables → Make predictions

data quality is critical because they do not rely on statistical theory and only learn patterns from data

HL7 FHIR

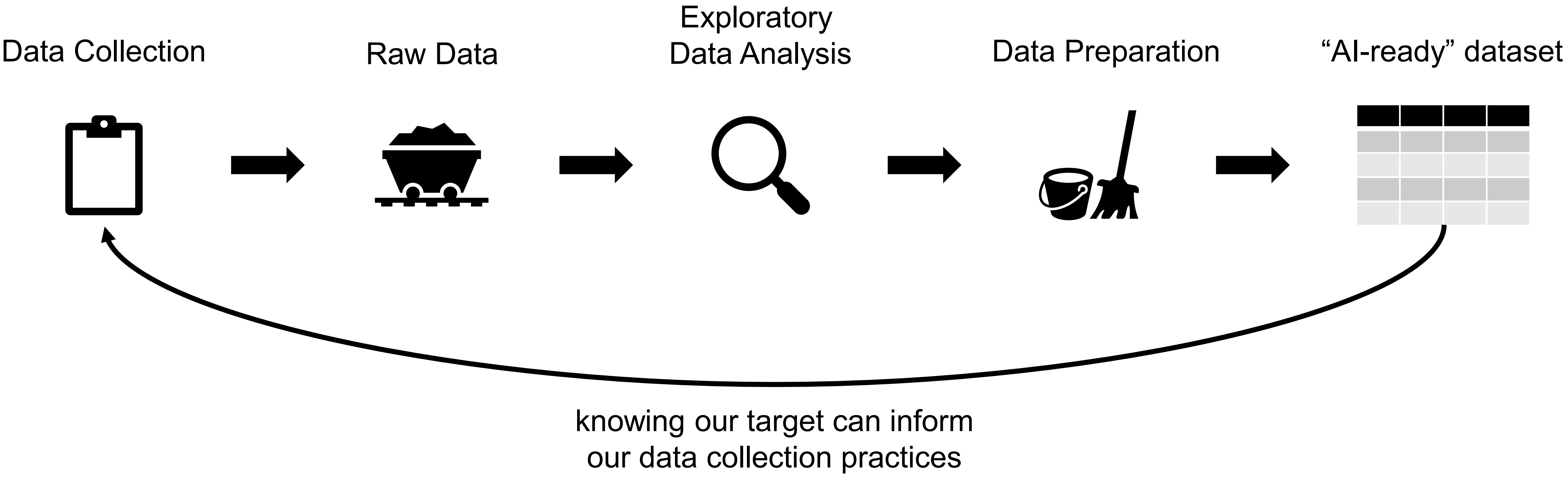
NIH CDE Repository

LOINC from Regenstrief

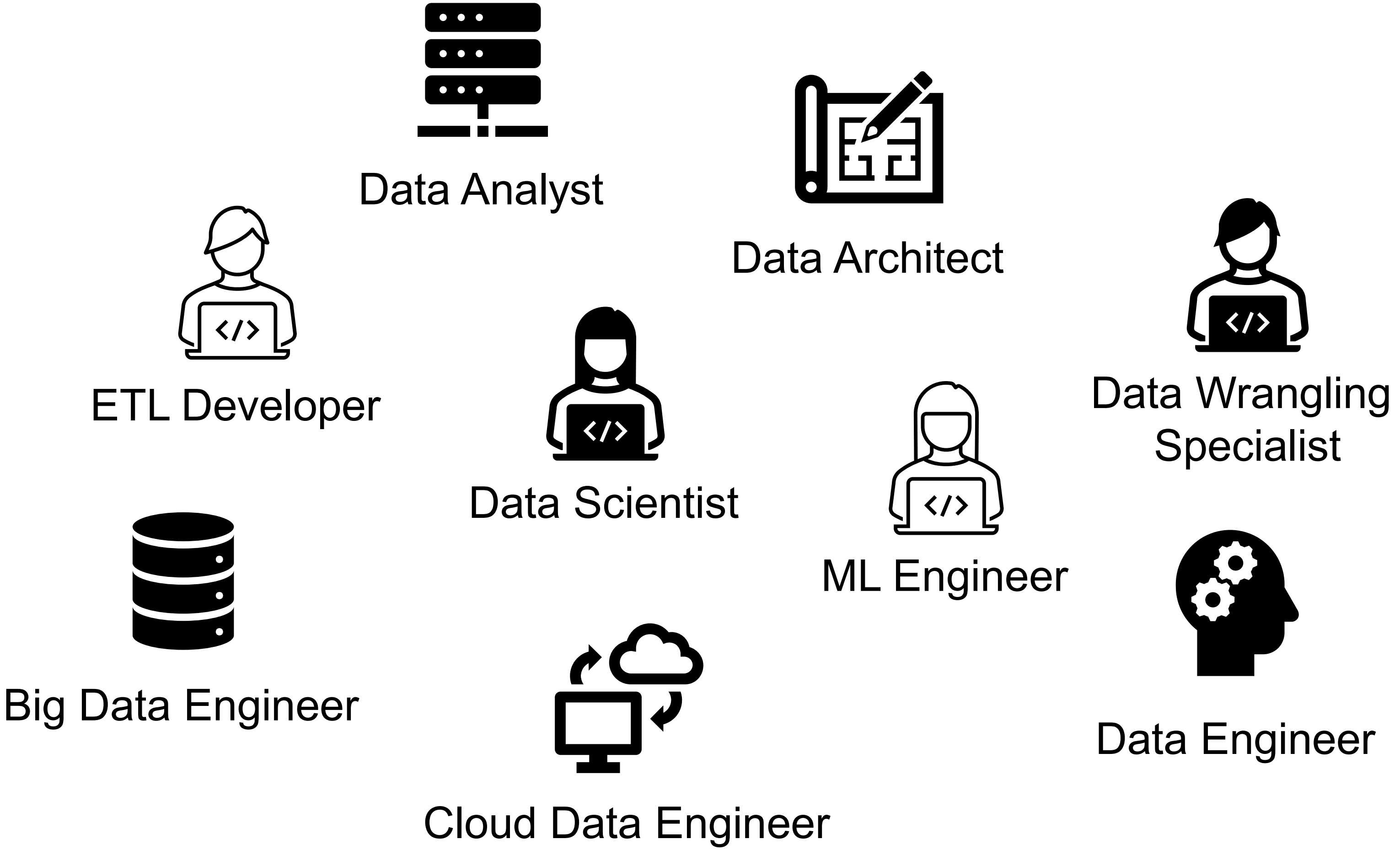
NIH NATIONAL CANCER INSTITUTE EVS Explore

the need for large datasets, necessitating linking across datasets, requires consistency in semantics and data standards

Understanding the requirements for AI-ready data informs data collection practices



The importance of data preparation has led to a bevy of roles – this is a specialized skill set that can/should be on grant applications

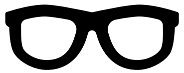


Visualizing Raw Data

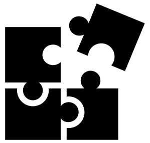
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Data Exploration

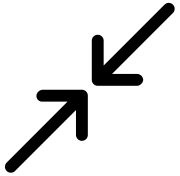
Data Exploration Goals



Understand what is in the data



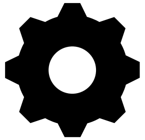
Assess data for project fit



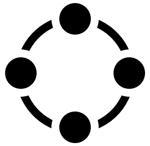
Reduce algorithm error

Data Exploration Steps

Inspection



Single variable



Variable relationships

Inspect data using simple descriptions and statistics

Inspection

See raw data table

```
df.head()
df.tail()
```

	Age	Blood Pressure	Cholesterol
0	63	137.0	140.0
1	76	144.0	217.0
2	53	96.0	288.0
3	39	121.0	168.0
4	67	112.0	281.0
5	32	111.0	195.0
6	45	128.0	302.0
7	63	110.0	238.0
8	43	112.0	154.0
9	47	117.0	272.0

Check Data Types

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 1000 non-null   int64
 1   Gender              1000 non-null   object
 2   Blood Pressure      970 non-null    float64
 3   Cholesterol         970 non-null    float64
 4   Diabetes            1000 non-null   int64
 5   Smoking             1000 non-null   int64
 6   Exercise Frequency  970 non-null    object
 7   BMI                 970 non-null    float64
 8   Family History      1000 non-null   int64
 9   Heart Disease       1000 non-null   int64
dtypes: float64(3), int64(5), object(2)
memory usage: 78.2+ KB
```

Compute summary statistics

```
df.describe()
```

	Age	Blood Pressure	Cholesterol
count	1000.000000	970.000000	970.000000
mean	52.852000	120.970103	202.490722
std	16.069796	18.164997	53.375961
min	25.000000	67.000000	33.000000
25%	39.750000	110.000000	166.000000
50%	53.000000	120.000000	201.500000
75%	66.000000	130.000000	237.000000
max	80.000000	252.000000	420.000000

Data Exploration

Data Exploration Goals

Understand what is in the data

Assess data for project fit

Reduce algorithm error

Data Exploration Steps

Inspection

Single variable

Variable relationships

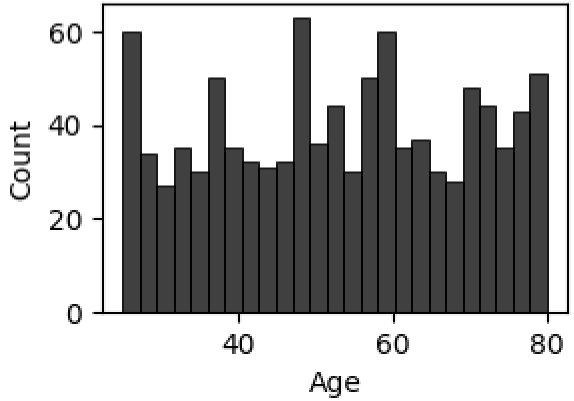
Single variable distributions inform downstream cleaning steps

Single variable

Numeric Data

Visualize distributions with histograms and boxplots

`sns.histplot()`



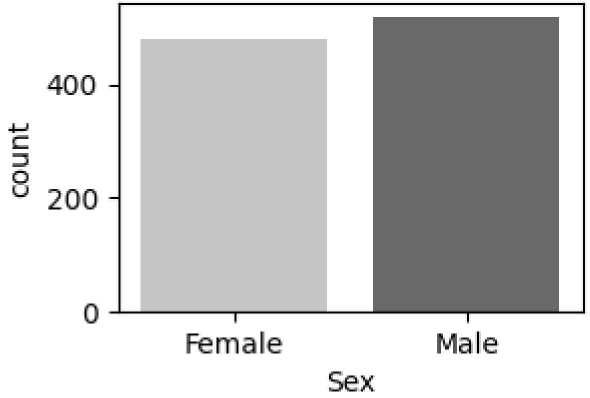
Visualize trends for timeseries data

`plt.plot()`

Categorical or Binary Data

Visualize distributions with bar charts and value counts

`sns.countplot()`



Data Exploration

Data Exploration Goals

Understand what is in the data

Assess data for project fit

Reduce algorithm error

Data Exploration Steps

Inspection

Single variable

Variable relationships

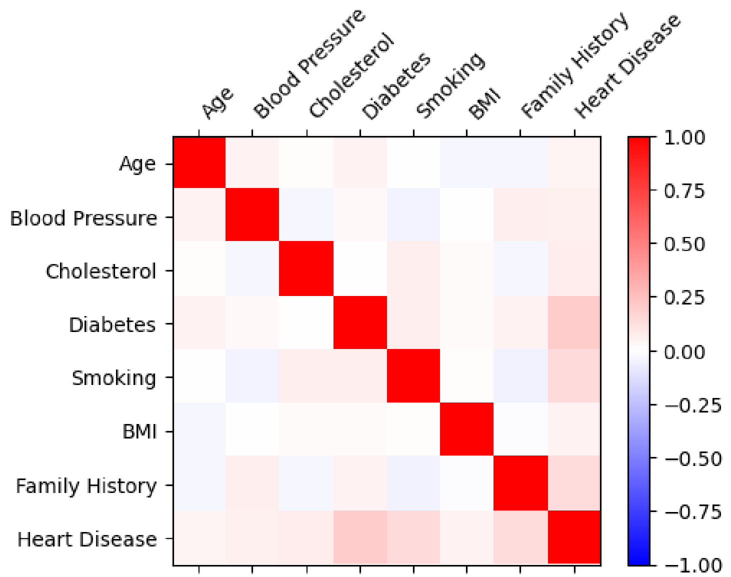
Examining variable relationships inform feature selection

Variable relationships

Numeric Data

See variable relationships with scatterplots and correlation matrices

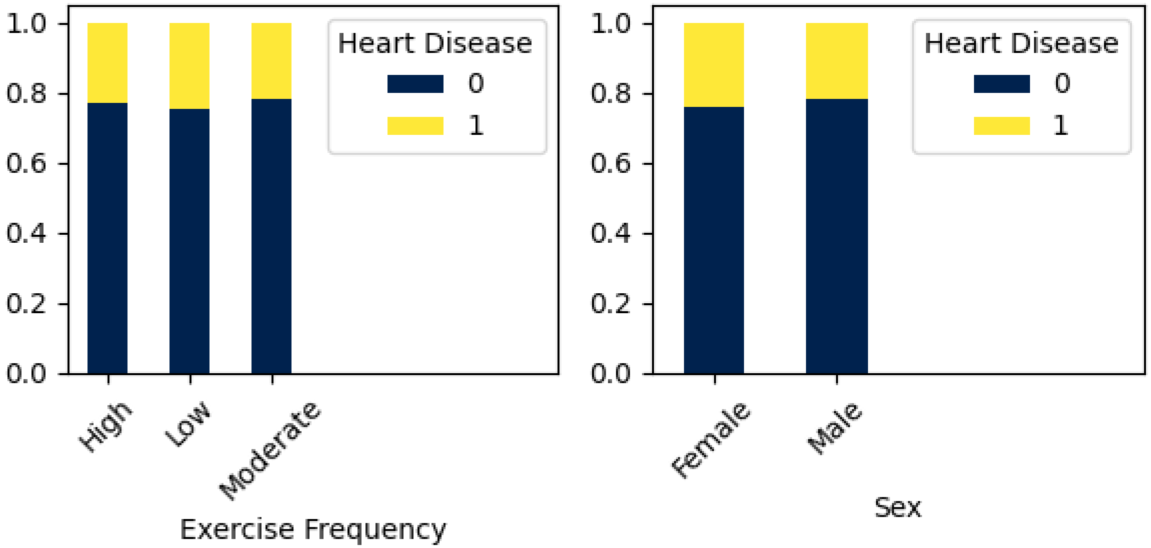
```
plt.matshow(df.corr())
```



Categorical or Binary Data

See variable relationships with cross-tabulation tables

```
pd.crosstab()
```



Data Exploration

Data Exploration Goals

Understand what is in the data

Assess data for project fit

Reduce algorithm error

Data Exploration Steps

Inspection

Single variable

Variable relationships

Slido Poll

Which of the following plot types is/are most appropriate for visualizing the distribution of a **categorical variable**?

a) Bar plot

b) Histogram

c) Box plot

d) Pie chart

e) Both A and D

Data Cleaning 101

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Data Cleaning

Data Cleaning Goals

- Meet baseline data requirements for model training
- Ensure data quality to reduce model error
- Enforce data format and model fit
- Enable data interoperability

Data Cleaning Steps

- Deal with Missing Data
- Handle Outliers
- Encode Non-numeric Variables

Understanding missing data

Deal with Missing Data

What is Missing Data?

CustomerID	Age	Income	City
001	28	NaN	London
002	NULL	75000	Paris
003	45	82000	""

Why Do We Have Missing Data?

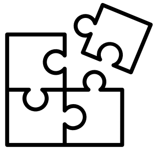
Corrupt Data

Human Error

Participant chose not to respond

Understanding missing data

Deal with Missing Data



Common representation of Missing Data

- NaN (Not a Number)
- NULL or None
- Empty String (“”)
- Special Indicators like 9999, -99
- Blank or Space

Types of Missing Data

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

Types of Missing Data

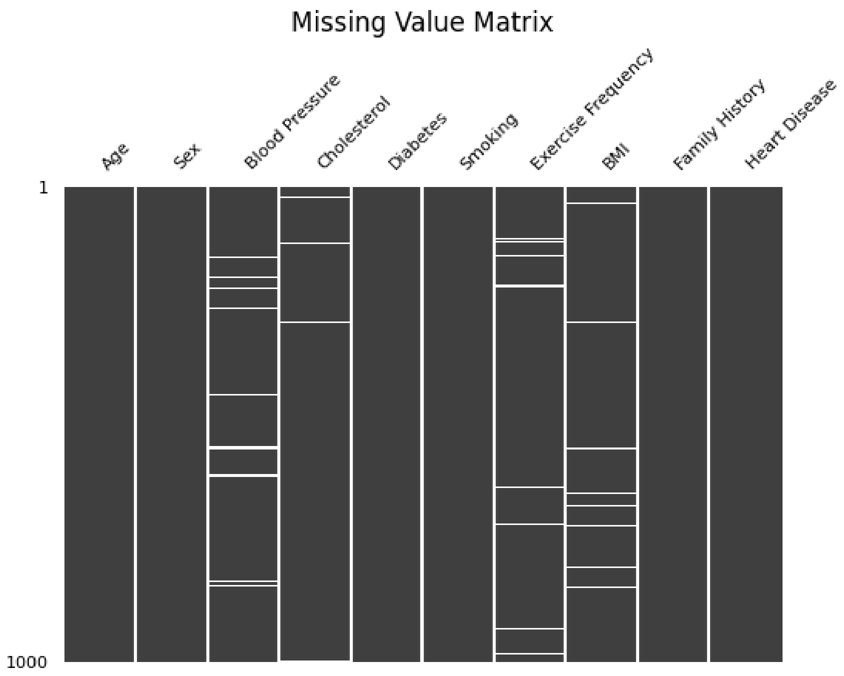
	Missing Completely at Random (MCAR)	Missing at Random (MAR)	Missing Not at Random (MNAR)
Problem	The missingness of data is not related to any other variable in the dataset. It is just random	The missingness of a variable is related to some other variables in the dataset but not the variable itself	The missingness of a variable is related to the variable itself
Example	A participant accidentally skipped cholesterol test due to lab error	Cholesterol data is missing more often in younger patients with no family history	Patients with very high cholesterol tend to hide results due to stigma

Understanding missing data

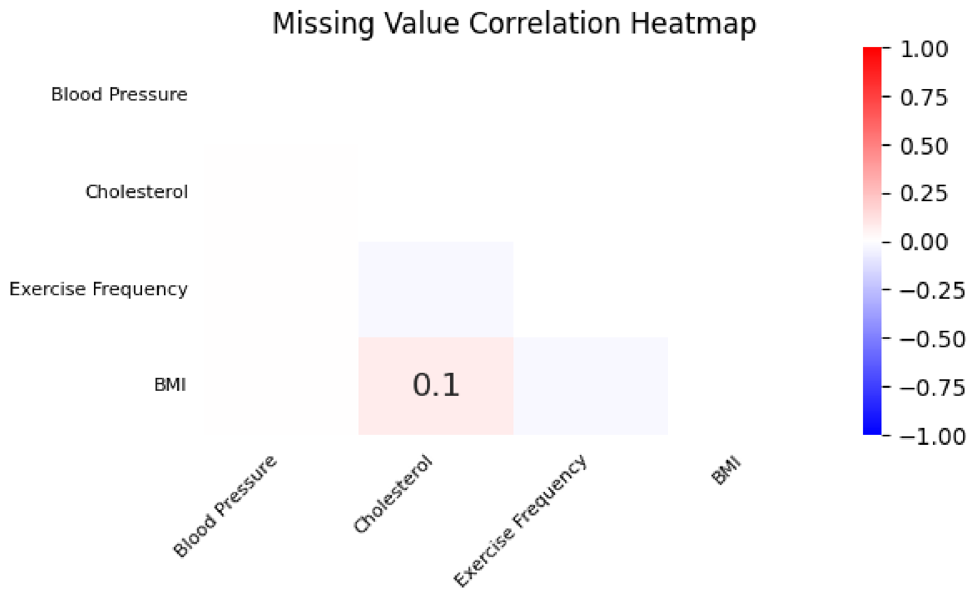
Deal with Missing Data

Visualize missing values pattern

```
msno.matrix(raw_data,ax=ax,sparkline=False)
```




```
msno.heatmap(raw_data, cmap='bwr',ax=ax)
```



Types of Missing Data

Deal with Missing Data

	Missing Completely at Random (MCAR)	Missing at Random (MAR)	Missing Not at Random (MNAR)
Problem	The missingness of data is not related to any other variable in the dataset. It is just random	The missingness of a variable is related to some other variables in the dataset but not the variable itself	The missingness of a variable is related to the variable itself
Example	A participant accidentally skipped cholesterol test due to lab error	Cholesterol data is missing more often in younger patients with no family history	Patients with very high cholesterol tend to hide results due to stigma
Solution	• Deletion • Simple Imputation	• Multiple Imputation • Predictive Imputation	• Advanced Model Based Imputation • Use of Proxy Variables • Sensitivity Analysis

Techniques for Handling Missing Data

Deal with Missing Data

Deletion

Imputation

Advanced Techniques

Techniques for Handling Missing Data

Deletion

- This is the simplest method, which involves deleting the records with missing values.
- However, it's only advisable when the data is Missing Completely at Random (MCAR) and the missing data is a small fraction of the total dataset.

Row-wise Deletion (or "Listwise Deletion")

Action

Removes entire observations (rows) that contain any missing values

```
df.dropna(inplace=True)
```

Best for

- Data is Missing Completely at Random (MCAR)
- Missing data represents a small percentage of the dataset

Tradeoffs

- **Pro:** Quick implementation
- **Con:** Reduces sample size, potentially decreasing statistical power
- **Con:** Can skew results (e.g., if missing values are related to specific groups)

Column-wise Deletion (or "Variable Deletion")

Action

Removes features (columns) with excessive missing values

```
threshold = 50
df_column_drop = df.dropna(axis=1,
thresh=len(raw_data) * (threshold / 100))
```

Best for

- When specific variables have high percentages of missing data
- When those variables aren't critical to your analysis

Tradeoffs

- **Pro:** Maintains the full observation count
- **Con:** Completely loses information from deleted variables
- **Con:** May eliminate potentially important predictors from the model

Techniques for Handling Missing Data

Deal with Missing Data

Deletion

Imputation

Advanced Techniques

Techniques for Handling Missing Data

Imputation

- Imputation is the process of substituting missing data with substituted values.
- There are many imputation methods for replacing the missing values.

Simple Imputation Techniques

```
df["var"].fillna(df["var"].median())
```

Pros

- Easy and fast to implement
- Preserves variable distributions

Cons

- Ignores feature relationships

Predictive Imputation Techniques

```
df[df["var"].isnull()]=
        rf_regressor.predict(df[df["var"].isna()])
```

Pros

- makes use of correlation information between variables

Cons

- Computationally expensive to implement

Techniques for Handling Missing Data

Deal with Missing Data

Deletion

Imputation

Advanced Techniques

Techniques for Handling Missing Data

Advanced Techniques

Advanced Imputation Techniques

```
df[“var”]=
        knn_impuer.fit_transform(df[“var”])
```

Pros
- makes use of correlation information between variables

Cons
- Computationally expensive to implement

Proxy Variables

```
df[“proxy”] =
      pd.read_csv(‘related_dataset.csv’)
```

Pros
- make use of a more widely available metric with more complete data

Cons
- proxy variable may not capture important correlations of the original variable
- It may not fully represent the intention- stakeholders input on this to handle properly

Techniques for Handling Missing Data

Deal with Missing Data

Deletion

Imputation

Advanced Techniques

Slido Poll

You're analyzing survey data from a health study. You notice that responses to the question *"Do you consume alcohol?"* are missing more frequently among participants over the age of 60. However, within each age group, the missingness appears random.

Based on this information, what is the most likely type of missingness?

a) Missing Completely at Random (MCAR)

b) Missing at Random (MAR)

c) Missing Not at Random (MNAR)

d) Not Missing — this is expected behavior

Techniques for Handling Missing Data

Deal with Missing Data

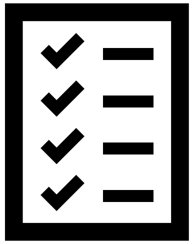
Deletion

Imputation

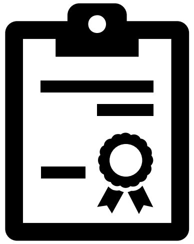
Advanced Techniques

Data Cleaning

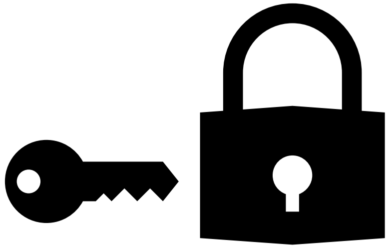
Data Cleaning Goals



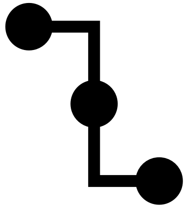
Meet baseline data requirements for model training



Ensure data quality to reduce model error



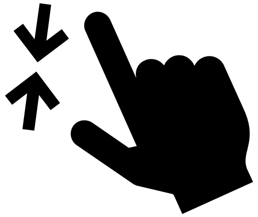
Enforce data format and model fit



Enable data interoperability

Data Cleaning Steps

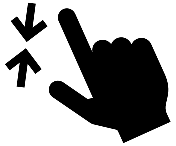
Deal with Missing Data



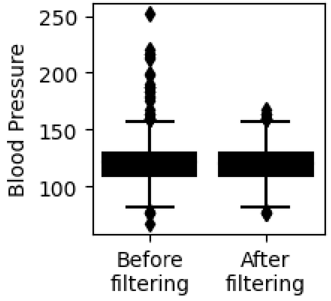
Handle Outliers

Encode Non-numeric Variables

Strategies for handling outlier data

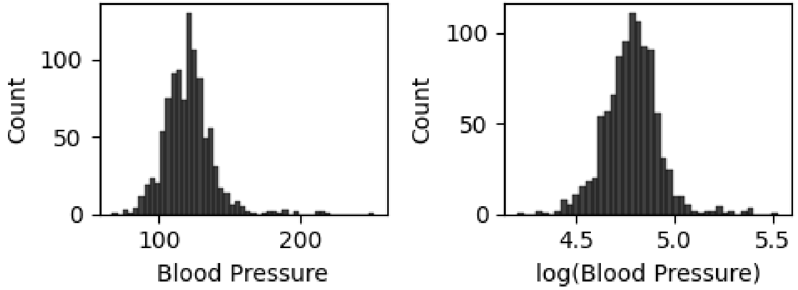


Remove outliers beyond set bounds



- **Pro:** prevent model from over-fitting
- **Con:** may distort the distribution

Transform data to reduce outlier impact



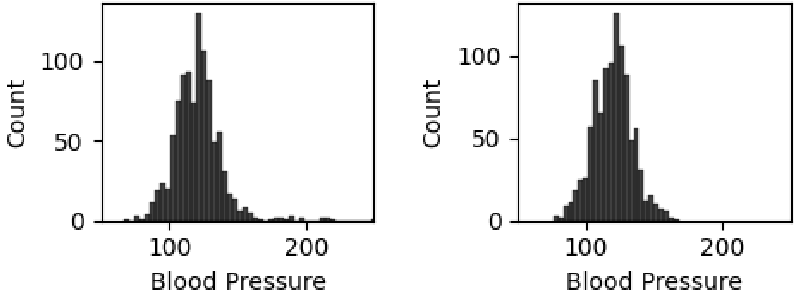
- **Pro:** improve data normality while retaining points
- **Con:** reduce variance and potentially biases data

Use models or scaling less sensitive to outliers

- Decision Trees
- Random Forest
- XGBoost
- AdaBoost
- Naive Bayes

  - **Pros:** preserves original data distribution
  - **Cons:** limits model choice

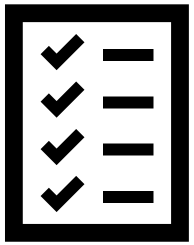
Impute or replace outliers



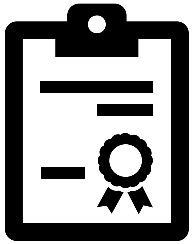
- **Pros:** preserve the features distribution
- **Cons:** lose information from outliers

Data Cleaning

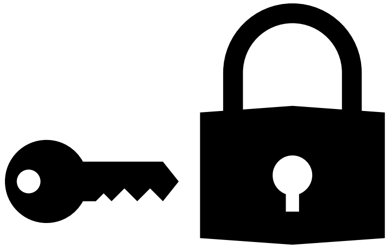
Data Cleaning Goals



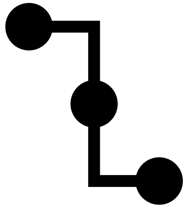
Meet baseline data requirements for model training



Ensure data quality to reduce model error



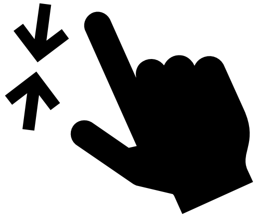
Enforce data format and model fit



Enable data interoperability

Data Cleaning Steps

Deal with Missing Data



Handle Outliers

Encode Non-numeric Variables

What are Non-Numeric Variables

Encode Non-numeric Variables

Non-numeric variables, also referred to as **categorical variables**, are data types that represent categories or distinct groups, rather than quantitative values. These variables are often used to capture qualitative attributes in a dataset.

Nominal Variables

- Blood type: A, B, AB, O
- Marital status: Single, Married, Divorced, Widowed
- Eye color: Brown, Blue, Green, Hazel

Ordinal Variables

- Education level: High School, Bachelor's, Master's, PhD
- Movie ratings: 1-star, 2-star, 3-star, 4-star, 5-star
- Economic status: Low-income, Middle-income, High-income

Binary Variables

- Yes/No responses
- True/False values
- Pass/Fail outcomes
- Employed/Unemployed status
- Presence/Absence of a condition

Turning categorical variables numeric is called encoding

One-hot Encoding

Sex
M
F
F
M

→

Male	Female
1	0
0	1
0	1
1	0

Pros
- good for data where order doesn't matter (nominal)

Cons
- adds many, sparse dimensions to data

Label Encoding

Sex
M
F
F
M

→

Sex
0
1
1
0

Pros
- good for data where order doesn't matter (nominal) and you have many values

Cons
- adds bias to models that assume numeric relationships

Ordinal Encoding

Exercise Frequency
Low
High
Moderate
Low

→

Exercise Frequency
0
2
1
0

Pros
- good for data where order matters (ordinal)
- more memory efficient than one-hot

Cons
- adds bias to models that assume numeric relationships

Data Cleaning

Data Cleaning Goals

Meet baseline data requirements for model training

Ensure data quality to reduce model error

Enforce data format and model fit

Enable data interoperability

Data Cleaning Steps

Deal with Missing Data

Handle Outliers

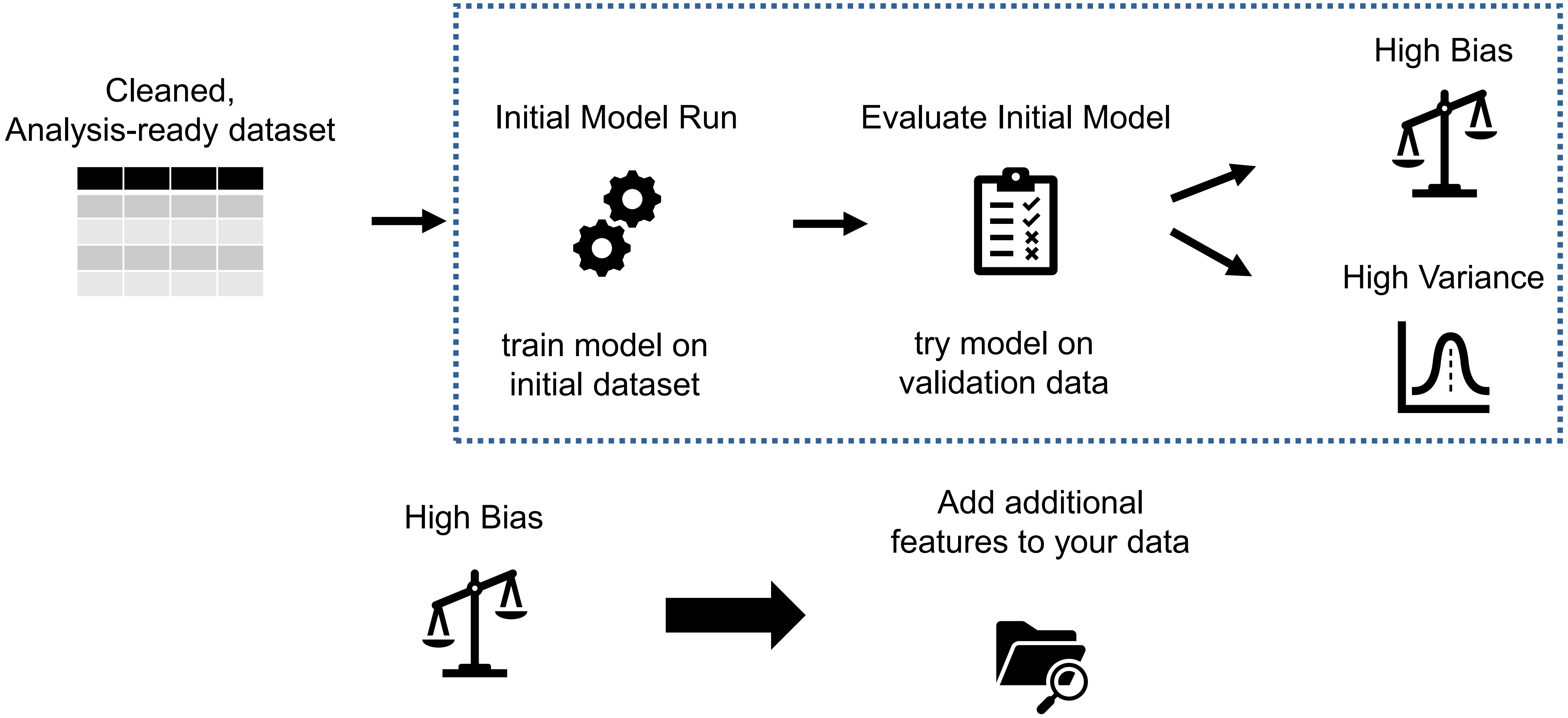
Encode Non-numeric Variables

Combining Data for Improving Model Performance

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

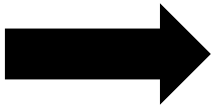
Testing if the dataset is sufficient for model development

to be discussed in depth in future Think-a-thons



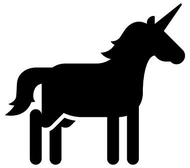
Domain expertise guides selection of additional data

High Bias



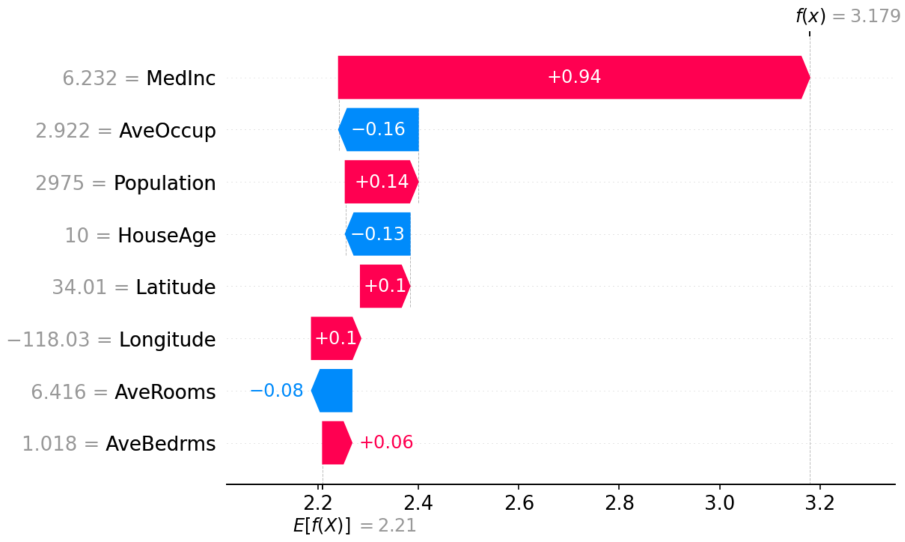
Add additional
features to your data

But, YOUR domain expertise can be a useful guide in deciding which datasets and what variables to include in the model to increase predictive power

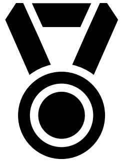


Also a place for
community input!

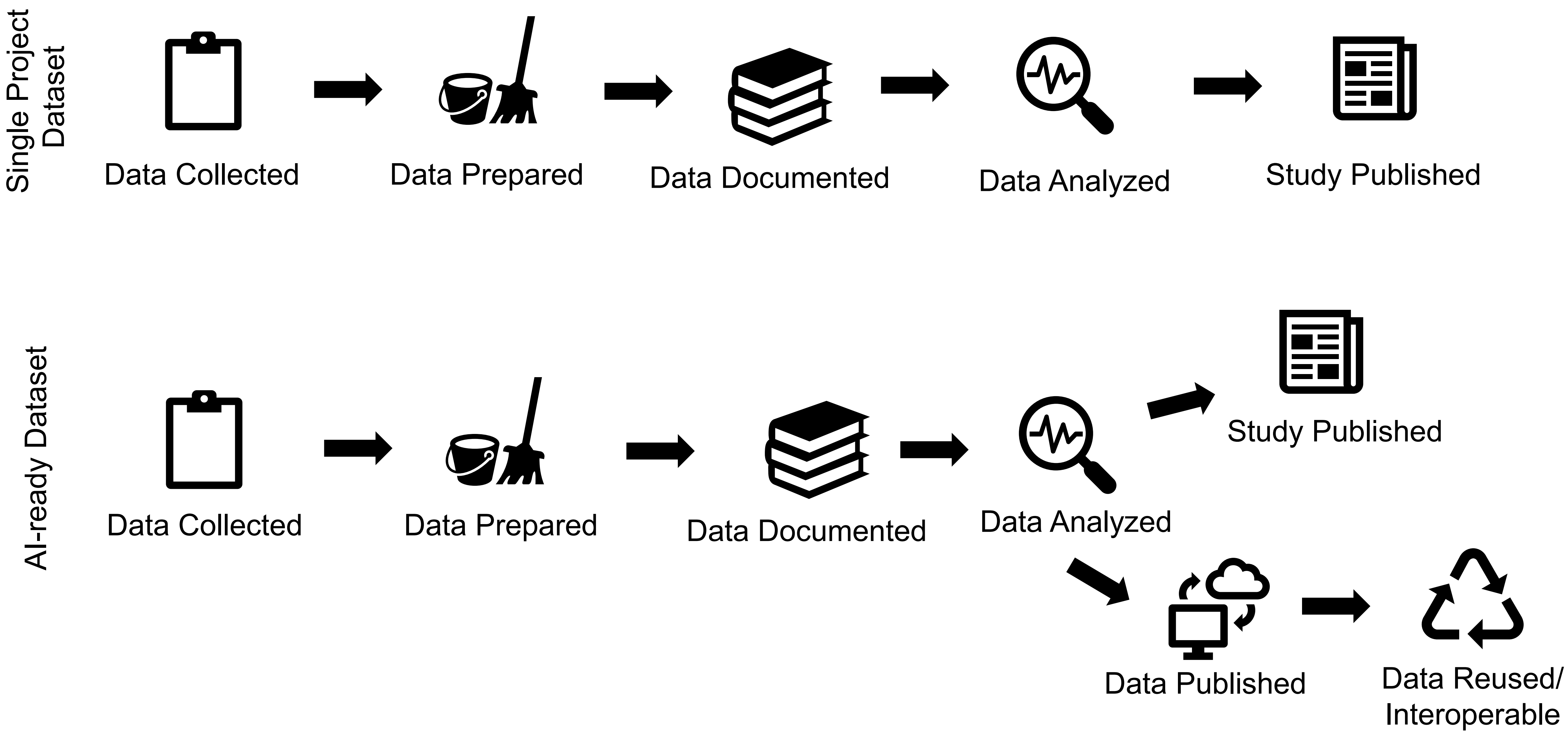
Technically, you could add in any and all data that is properly prepared and let the model distinguish useful features for prediction



Combining technical tests (like bias vs variance) with your domain expertise will create optimal models that also provide useful insights from analyzing feature importance



Data reusability relies on comprehensive data preparation



*We encourage all SCHARE members to create .csv files when uploading to the SCHARE Data Repository

Data Preparation

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Data Preparation

Data Preparation Goals

Enable data interoperability

Data Preparation Steps

Aligning Data Labels

Variable Transformation

Joining via Common Identifiers

Joining datasets requires common data labels

Aligning Data Labels

Dataset 1

Participant ID	Age
001	37
002	71
003	49
004	52

Dataset 2

Participant ID	Age at Enrollment
005	48
006	24
007	83
008	55

Dataset 2

Participant ID	Age at Enrollment
005	48
006	24
007	83
008	55

Dataset 2

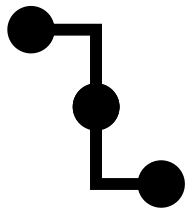
Participant ID	Age
005	48
006	24
007	83
008	55

Merge with Dataset 1

Participant ID	Age
001	37
002	71
003	49
004	52
005	48
006	24
007	83
008	55

Data Preparation

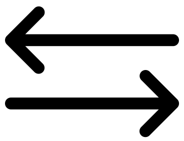
Data Preparation Goals



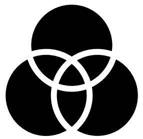
Enable data interoperability

Data Preparation Steps

Aligning Data Labels



Variable Transformation



Joining via Common Identifiers

Concatenating datasets requires variable alignment

Dataset 1

Participant ID	Age at Enrollment	BPM
001	37	62
002	71	73
003	49	52
004	52	61

Dataset 2

Participant ID	Date of Birth	BPM
005	1972	63
006	1984	54
007	2001	77
008	1992	61

Concatenating these sets requires transformation of the age variable

Dataset 2

Date of Birth
1972
1984
2001
1992

Dataset 2

Age
53
41
24
33

Variable Transformation

Data Preparation

Data Preparation Goals

Enable data interoperability

Data Preparation Steps

Aligning Data Labels

Variable Transformation

Joining via Common Identifiers

Joining datasets with common subjects requires identifiers

Dataset 1

ZIP	Population
01581	20000
99701	9000
27292	15000
33145	23000

Dataset 2

Town Name	Area (sq. mi)
Westborough, MA	8
Fairbanks, AK	30
Ashburn, VA	3
Miami, FL	9

Joining via
Common Identifiers

Joining these datasets requires
transformation to a common identifier,
followed by a join

Dataset 2

Town Name
Westborough, MA
Fairbanks, AK
Ashburn, VA
Miami, FL

Dataset 2

ZIP
01581
99701
27292
33145

Joined dataset

ZIP	Population	Area (sq. mi)
01581	20000	8
99701	9000	30
27292	15000	3
33145	23000	9

Data Preparation

Data Preparation Goals

Enable data
interoperability

Data Preparation Steps

Aligning Data Labels

Variable Transformation

Joining via Common Identifiers

Working with Small Sample Sizes in Health Research

The Challenge

Small samples are common when studying:

- Underrepresented populations
- Rare health conditions
- Specific demographic intersections
- Areas with data access limitations

Why This Matters

- ML algorithms typically expect **large datasets**
- Small samples can lead to **unreliable models**
- Critical for **accurate representation** of all communities

Key Risks

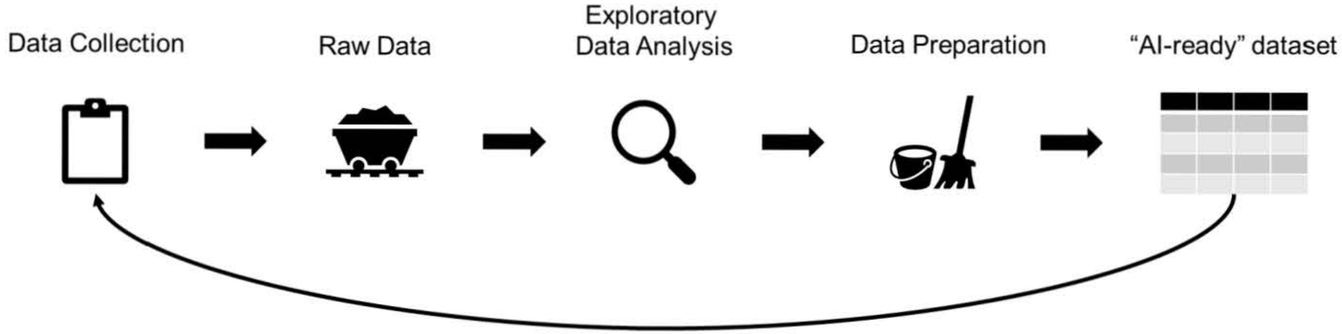
- **Overfitting:** Models learn noise, not patterns
- **Limited Generalizability:** Findings don't transfer to broader population
- **High Variance:** Results change dramatically with small data shifts
- **Missed Insights:** Important health differences go undetected

How to Prepare Small Sample Sizes

Strategy	Why it Helps	What is solves
Feature Selection	Reduce dimensionality and noise	Minimizes overfitting and simplifies models with limited data
Robust Imputation	Avoid losing rows—Missing at Random (MAR)-aware techniques preserve data	Preserves dataset size and reduces error from missing data
Data Augmentation	Generate synthetic samples (e.g., SMOTE) for rare subgroups	Mitigates class imbalance and enhances learning from small/rare classes
Cross-validation	Helps get stable estimates from limited data	Reduces variance in model evaluation; more reliable performance metrics
External data merging	Use public datasets to enrich limited features	Expands feature space and improves model generalization

Data Cleaning Summary

Data Preparation is Essential for AI/ML Analyses



Goals of Data Preparation

- Meet baseline data requirements for model training
- Ensure data quality to reduce model error
- Enforce data format and model fit
- Enable data interoperability

Methods for Raw Data Exploration

- Inspection
- Single variable
- Variable relationships

Methods for Data Preparation

- Deal with Missing Data
  - Deletion
  - Imputation
  - Advanced Techniques
- Handle Outliers
- Encode Non-numeric Variables

Considerations for small datasets

Slido Poll

Which of the following is **NOT** a common technique for handling outliers in a dataset?

a) Removing outliers based on domain knowledge or statistical thresholds

b) Applying transformations such as log or square root

c) Imputing outliers with mean or median values

d) Replacing outliers with random values from a normal distribution

e) Using robust models that are less sensitive to outliers (e.g., median-based regression)

Demo!

Data Preparation Notebook

1. Register for Terra (if you don’t have an account)
2. Put your Terra email address (usually gmail) in the form using the link
3. Access the workspace
4. Open the notebook corresponding with your last initial
5. Hit Run in Playground Mode

SCHARE

Thank you

Think-a-Thon poll

1. **Rate how useful this session was:**

- ☐ Very useful
- ☐ Useful
- ☐ Somewhat useful
- ☐ Not at all useful

Think-a-Thon poll

2. Rate the pace of the instruction for yourself:

- ☐ Too fast
- ☐ Adequate for me
- ☐ Too slow

Think-a-Thon poll

3. How likely will you participate in the next Think-a-Thon?

- ☐ Very interested, will definitely attend
- ☐ Interested, likely will attend
- ☐ Interested, but not available
- ☐ Not interested in attending any others

SCHARE

Next Think-a-Thons:

bit.ly/think-a-thons

Register for SCHARE:

https://bit.ly/registerschare

schare@mail.nih.gov