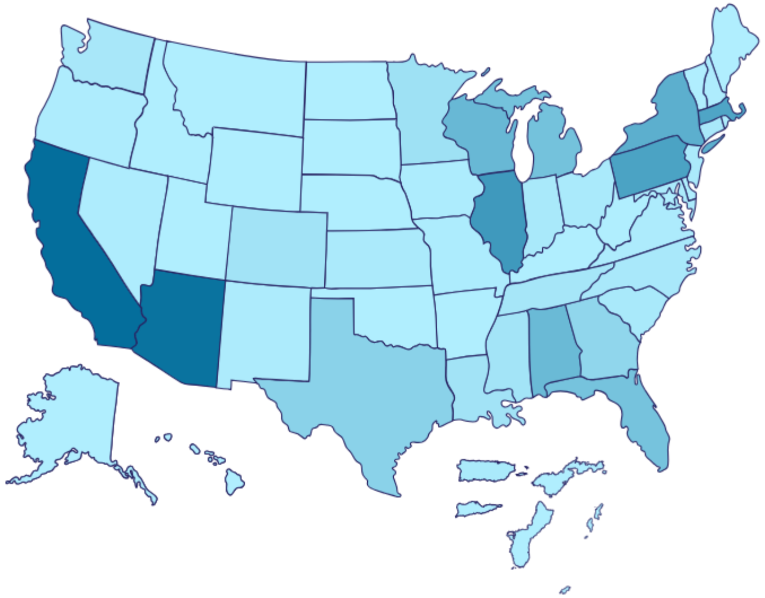
Getting Ready to Use the *All of Us* Researcher Workbench: Using Jupyter Notebooks with R

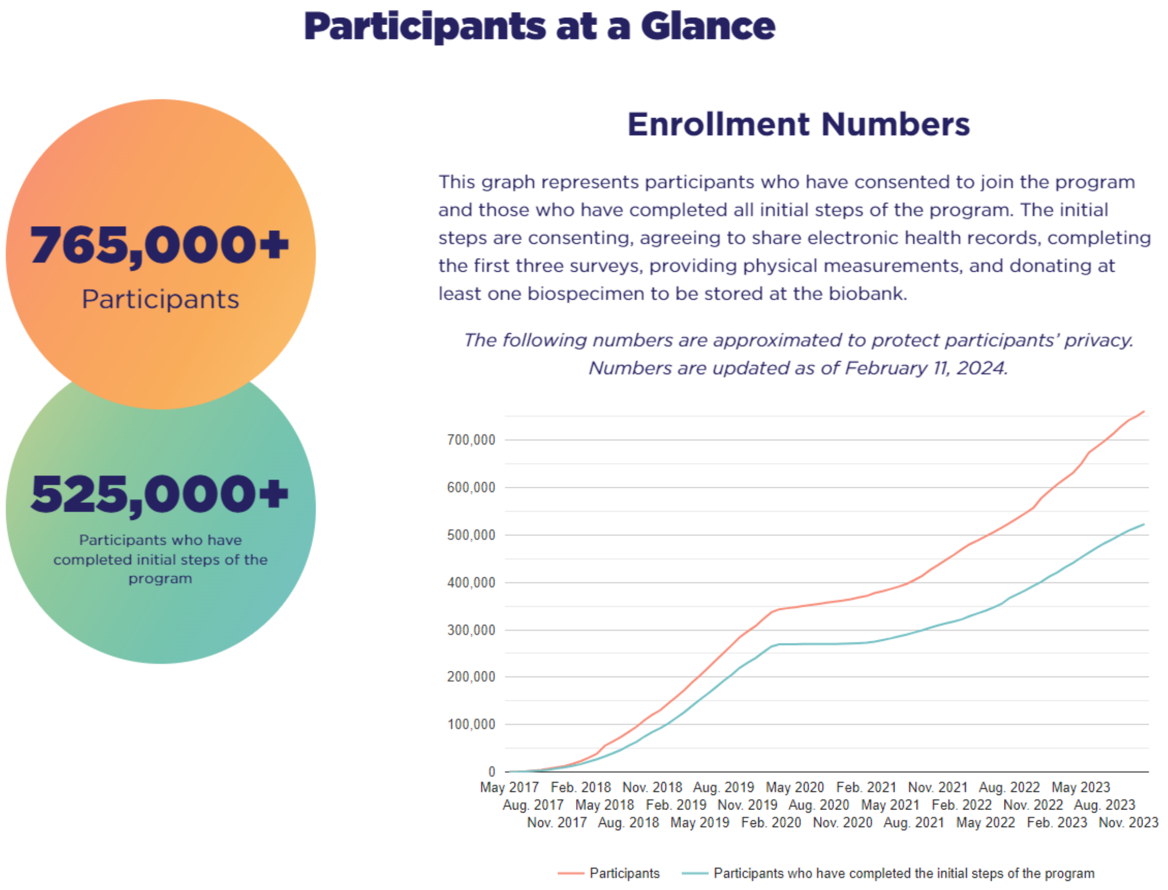
Learning Objectives

- *All of Us* Research Program background and description
- Terminology, concepts and structures
- Jupyter Notebooks and R

Who are the participants in All of Us?

1,163,00 people have registered





Participants at a Glance

765,000+

Participants

525,000+

Participants who have completed initial steps of the program

Enrollment Numbers

This graph represents participants who have consented to join the program and those who have completed all initial steps of the program. The initial steps are consenting, agreeing to share electronic health records, completing the first three surveys, providing physical measurements, and donating at least one biospecimen to be stored at the biobank.

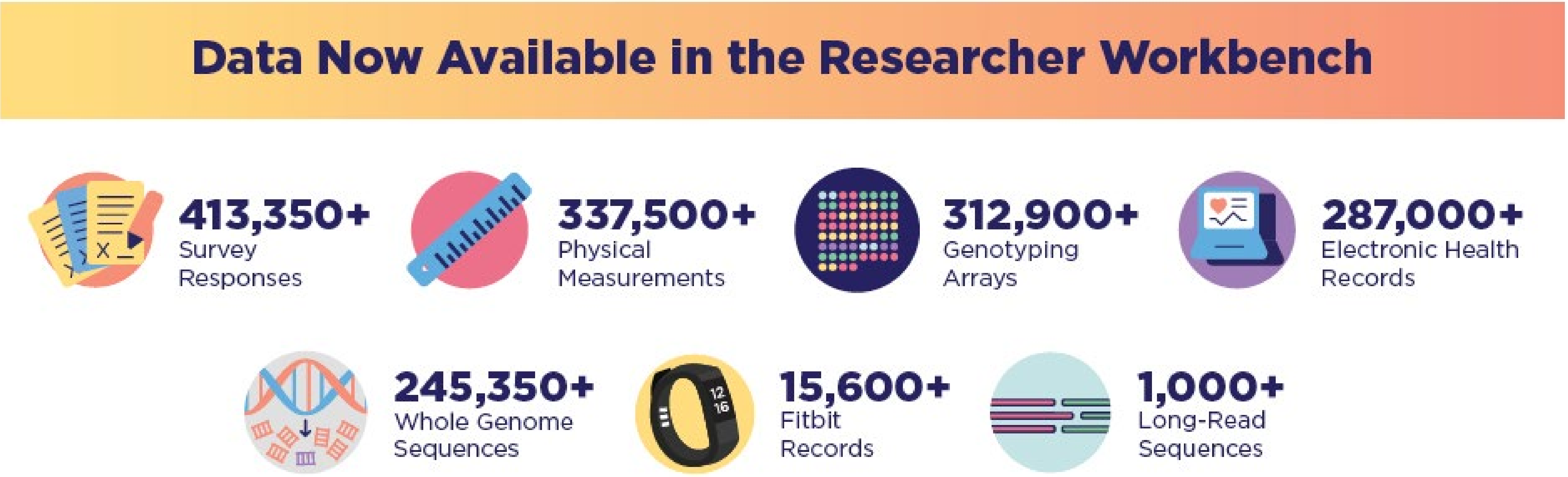
The following numbers are approximated to protect participants' privacy. Numbers are updated as of February 11, 2024.

Kinds of Access to the Datasets

There are three levels of access to the All of Us datasets. All data is stored in the “Research Hub” https://www.researchallofus.org/

1. Public Tier
2. Registered Tier
3. Controlled Tier

What kind of data is in All of Us?



Data Now Available in the Researcher Workbench

413,350+
Survey Responses

337,500+
Physical Measurements

312,900+
Genotyping Arrays

287,000+
Electronic Health Records

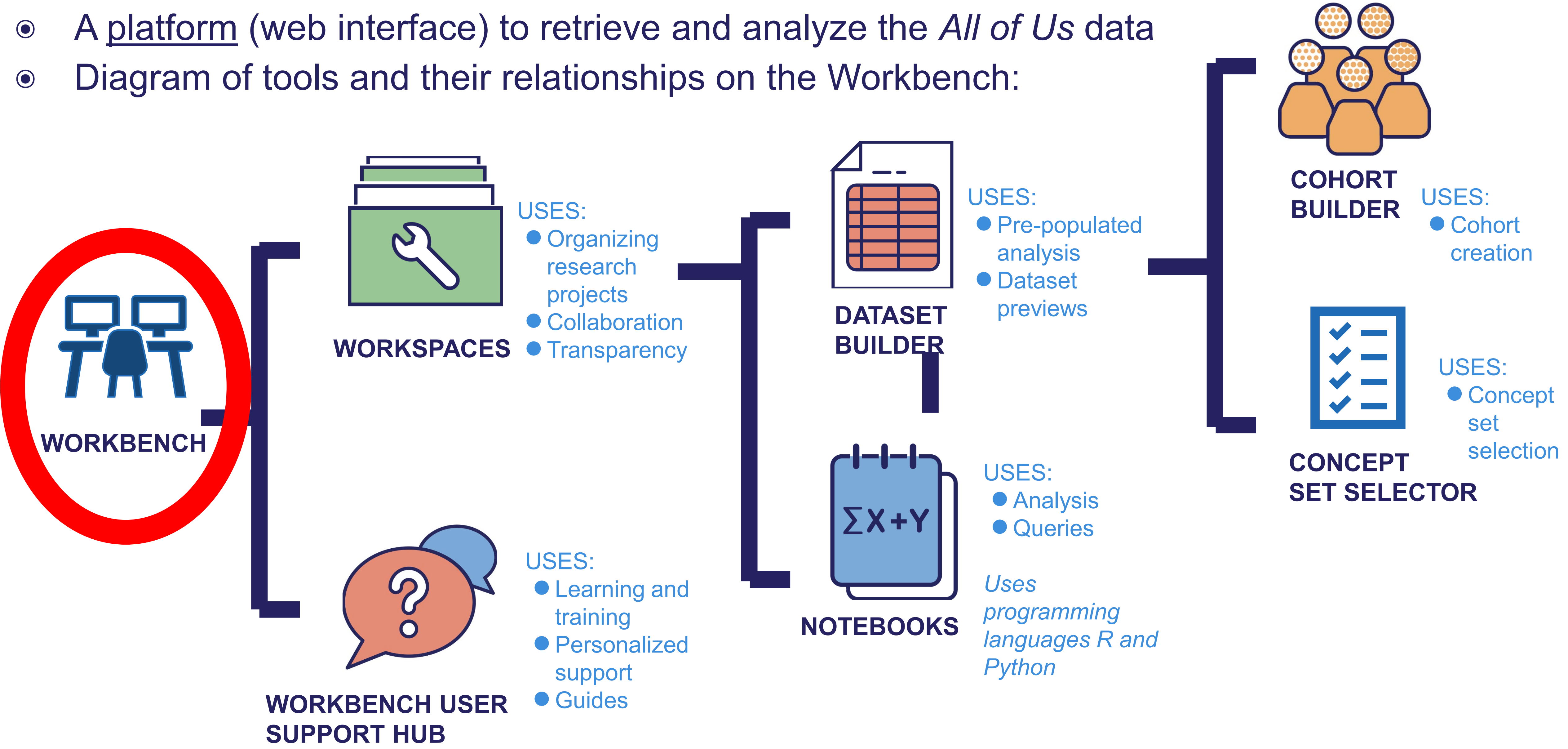
245,350+
Whole Genome Sequences

15,600+
Fitbit Records

1,000+
Long-Read Sequences

What is the *All of Us* Researcher Workbench?

- A platform (web interface) to retrieve and analyze the *All of Us* data
- Diagram of tools and their relationships on the Workbench:



WORKBENCH

WORKSPACES

USES:
- Organizing research projects
- Collaboration
- Transparency

WORKBENCH USER SUPPORT HUB

USES:
- Learning and training
- Personalized support
- Guides

DATASET BUILDER

USES:
- Pre-populated analysis
- Dataset previews

NOTEBOOKS

USES:
- Analysis
- Queries

Uses programming languages R and Python

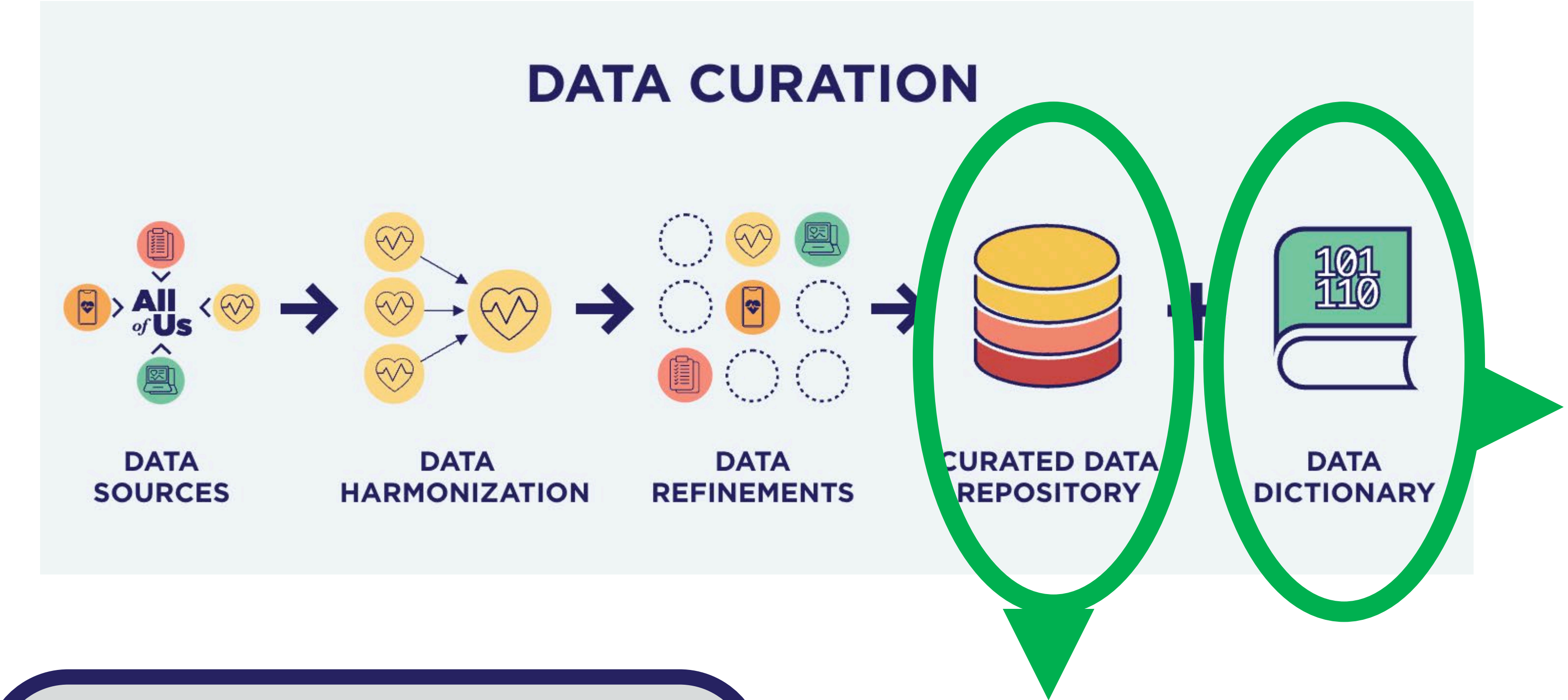
COHORT BUILDER

USES:
- Cohort creation

CONCEPT SET SELECTOR

USES:
- Concept set selection

Where is the data that is being accessed on the Jupyter Notebook?



The data dictionary is an Excel file:

- Spreadsheets of all the data types and explanations of them in more detail
- Link to it is under ‘Data Methods’ in in the Research Hub

The Curated Data Repository (CDR) is a BigQuery data warehouse:

- Cloud platform that stores the data
- SQL queries can retrieve data from the CDR one of two ways:
  - Using the Researcher Workbench tools to automatically generate SQL code in the Jupyter Notebook
  - Writing SQL code manually in the Jupyter Notebook

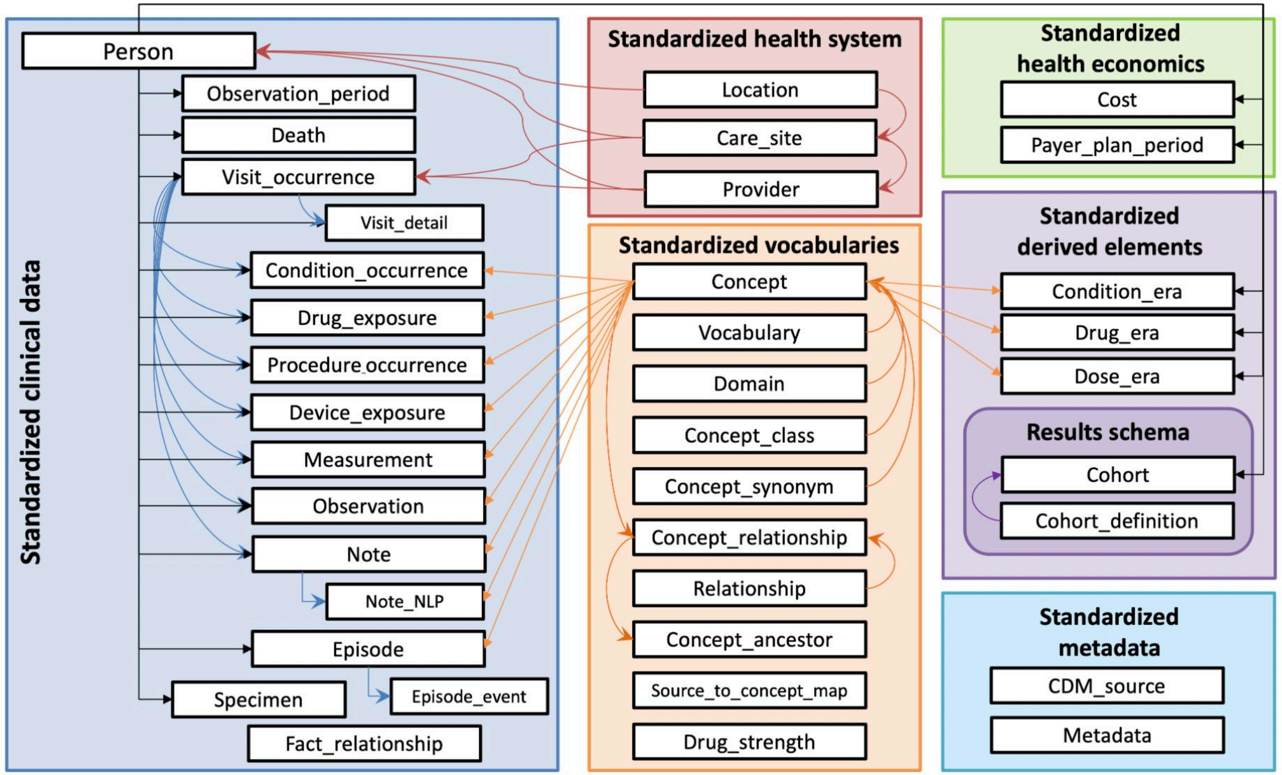
SQL

=

‘Structured Query Language”; SQL is a common programming language for using with relational databases

Where is the data that is being accessed on the Jupyter Notebook?

- The CDR has some data stored according to the OMOP common data model (CDM) from OHSDI (ohdsi.github.io/CommonDataModel/)
- More details can be found from the All of Us Office Hours 10/16/2020 "*AoU* use of OMOP for CDR" video https://youtu.be/jK11qAus8Q8



Standardized clinical data: Person, Observation_period, Death, Visit_occurrence, Visit_detail, Condition_occurrence, Drug_exposure, Procedure_occurrence, Device_exposure, Measurement, Observation, Note, Note_NLP, Episode, Episode_event, Specimen, Fact_relationship

Standardized health system: Location, Care_site, Provider

Standardized vocabularies: Concept, Vocabulary, Domain, Concept_class, Concept_synonym, Concept_relationship, Relationship, Concept_ancestor, Source_to_concept_map, Drug_strength

Standardized health economics: Cost, Payer_plan_period

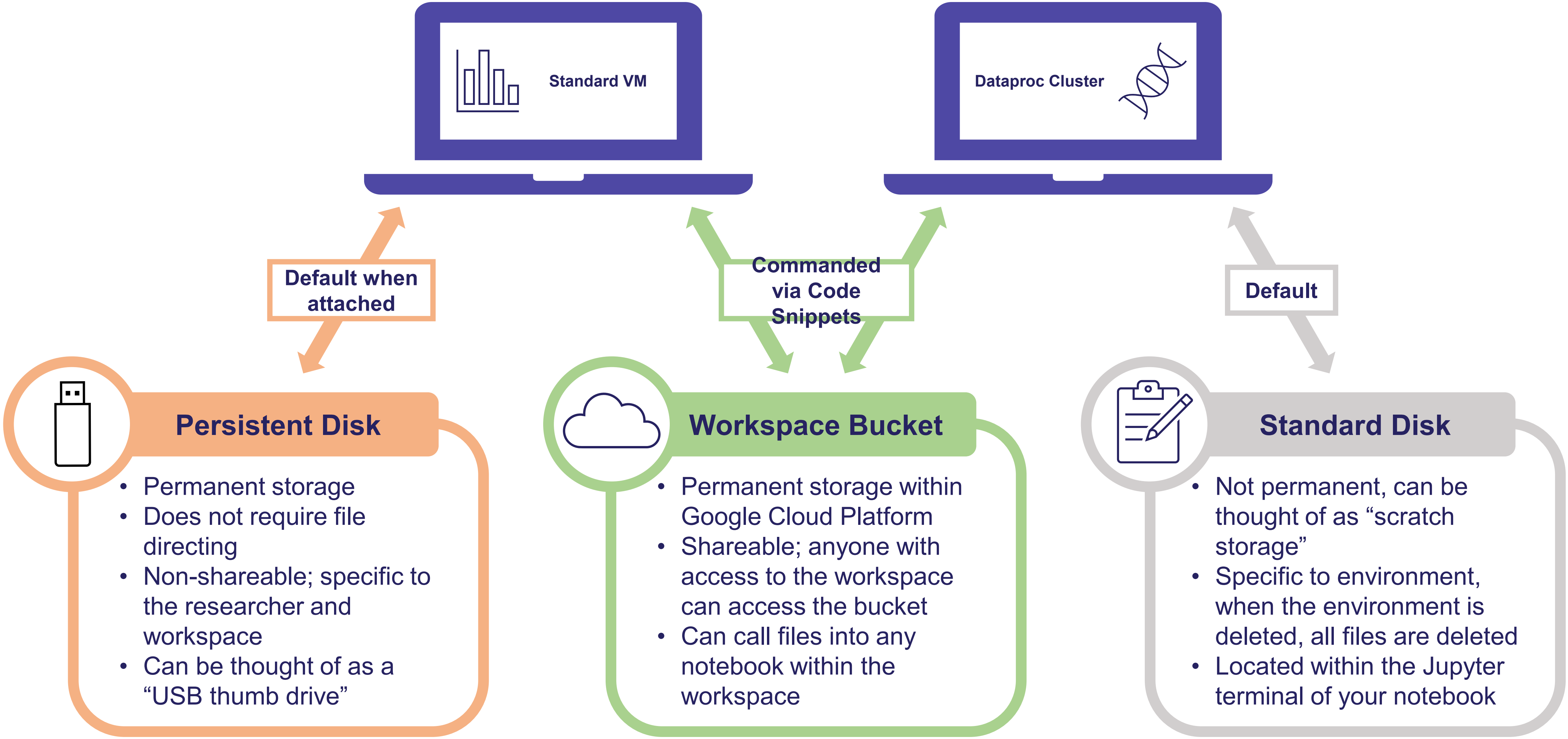
Standardized derived elements: Condition_era, Drug_era, Dose_era

Results schema: Cohort, Cohort_definition

Standardized metadata: CDM_source, Metadata

- we can retrieve our dataset of interest from CDR with code (**SQL queries**) by either using the Researcher Workbench web interface tools to generate this code **OR** by writing our own code in the Jupyter Notebook

Researcher Workbench Storage Options



Standard VM

Dataproc Cluster

Default when attached

Commanded via Code Snippets

Default

Persistent Disk

- Permanent storage
- Does not require file directing
- Non-shareable; specific to the researcher and workspace
- Can be thought of as a “USB thumb drive”

Workspace Bucket

- Permanent storage within Google Cloud Platform
- Shareable; anyone with access to the workspace can access the bucket
- Can call files into any notebook within the workspace

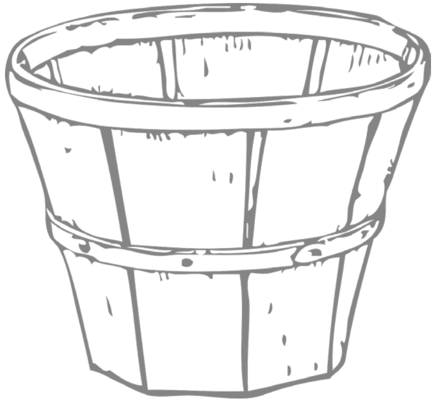
Standard Disk

- Not permanent, can be thought of as “scratch storage”
- Specific to environment, when the environment is deleted, all files are deleted
- Located within the Jupyter terminal of your notebook

What are Workspace Buckets, and what are they used for?

Workspace Bucket = **permanent cloud storage** available for your Workspace

- Each workspace has its own unique Uniform Resource Identifier (URI) which can be accessed from other Workspaces
- Saving files to the Workspace Bucket
  - Jupyter Notebooks are saved automatically
  - Other files must be saved manually
  - Files can be accessed across Jupyter Notebooks and Workspaces



- Take advantage of Workspace Buckets to save and retrieve your dataset
  - Easily retrieve a dataset without rewriting the SQL query or reformatting the data
  - May cost less to save and retrieve the data

I. Suggested Tutorials

A. Python
Udemy's 'Absolute Python Basics for Anyone' (video-oriented training)
https://www.udemy.com/course/absolute-python-basics-for-anyone/

Code Academy's 'Learn Python 3' (Reading and live coding-oriented training)
https://www.codecademy.com/learn/learn-python-3

Practical Computing for Biologists **Appendix 4** (reference sheets)
https://practicalcomputing.org/files/PCfB_Appendices.pdf

B. R
Udemy's 'R-Basics' (video-oriented training)
https://www.udemy.com/course/r-basics/

Code Academy's 'Learn R' (Reading and live coding-oriented training)
https://www.codecademy.com/learn/learn-r

'R Crash Course for Biologists' (book online)
https://github.com/ColauttiLab/RCrashCourse_Book/blob/master/ColauttiRCrashCourseNov22.pdf

'Computational Genomics with R' (book online)
https://compgenomr.github.io/book/

C. SQL
Khan Academy's 'Intro to SQL' (video and live coding-oriented training)
https://www.khanacademy.org/computing/computer-programming/sql

Code Academy's 'Learn SQL' (Reading and live coding-oriented training)
https://www.codecademy.com/learn/learn-sql