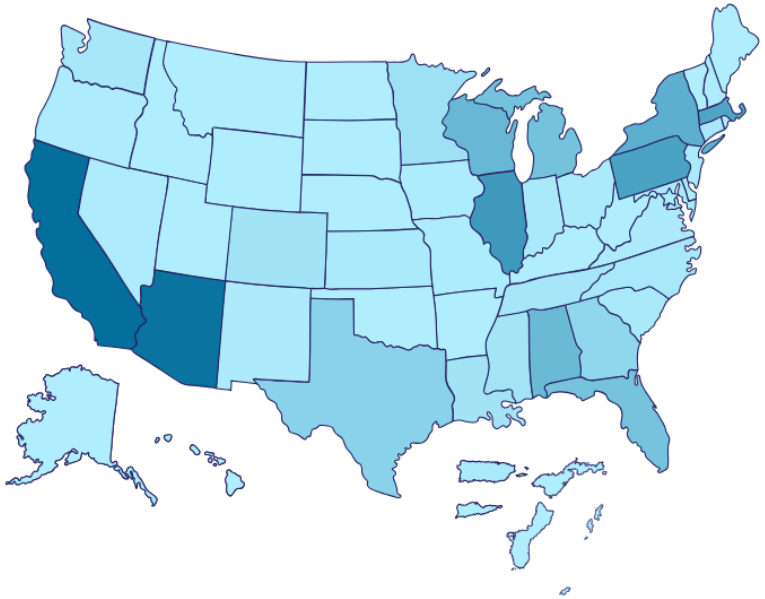
Getting Ready to Use the *All of Us* Researcher Workbench

Learning Objectives

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

Who are the participants in All of Us?

1,163,00 people have registered



Participants at a Glance

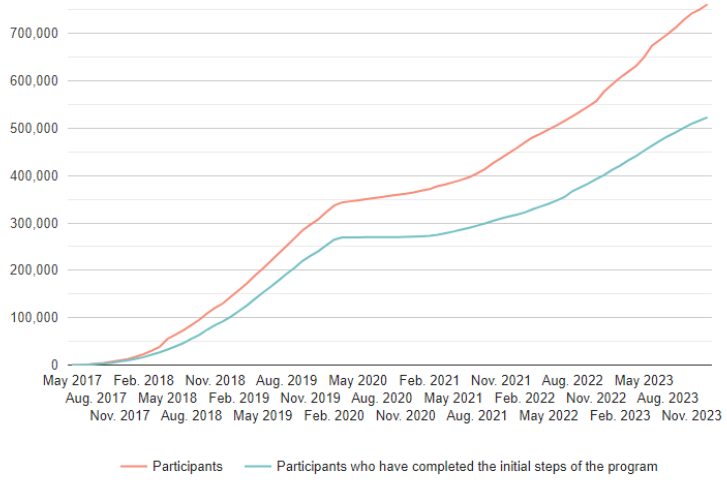
765,000+
Participants

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.



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

Kinds of Access to the Datasets

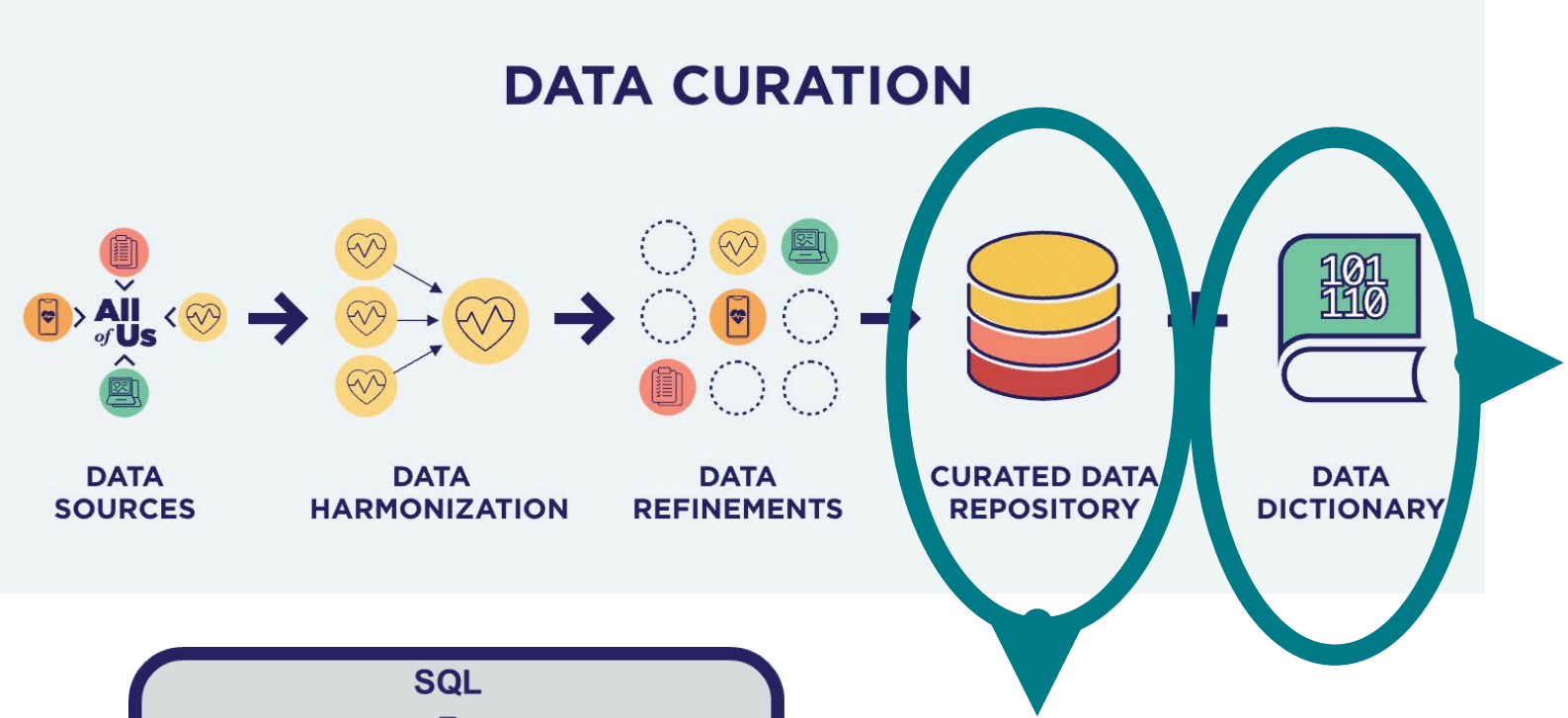
There are three levels of access to the All of Us datasets. All data is stored in the "Research Hub" https://www.researchallofus.org/

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

https://libguides.unm.edu/allofus

https://libguides.health.unm.edu/allofus

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 '<u>Data Methods</u>' in in the Research Hub

SQL = 'Structured Query Language"; SQL is a common programming language for using with relational databases

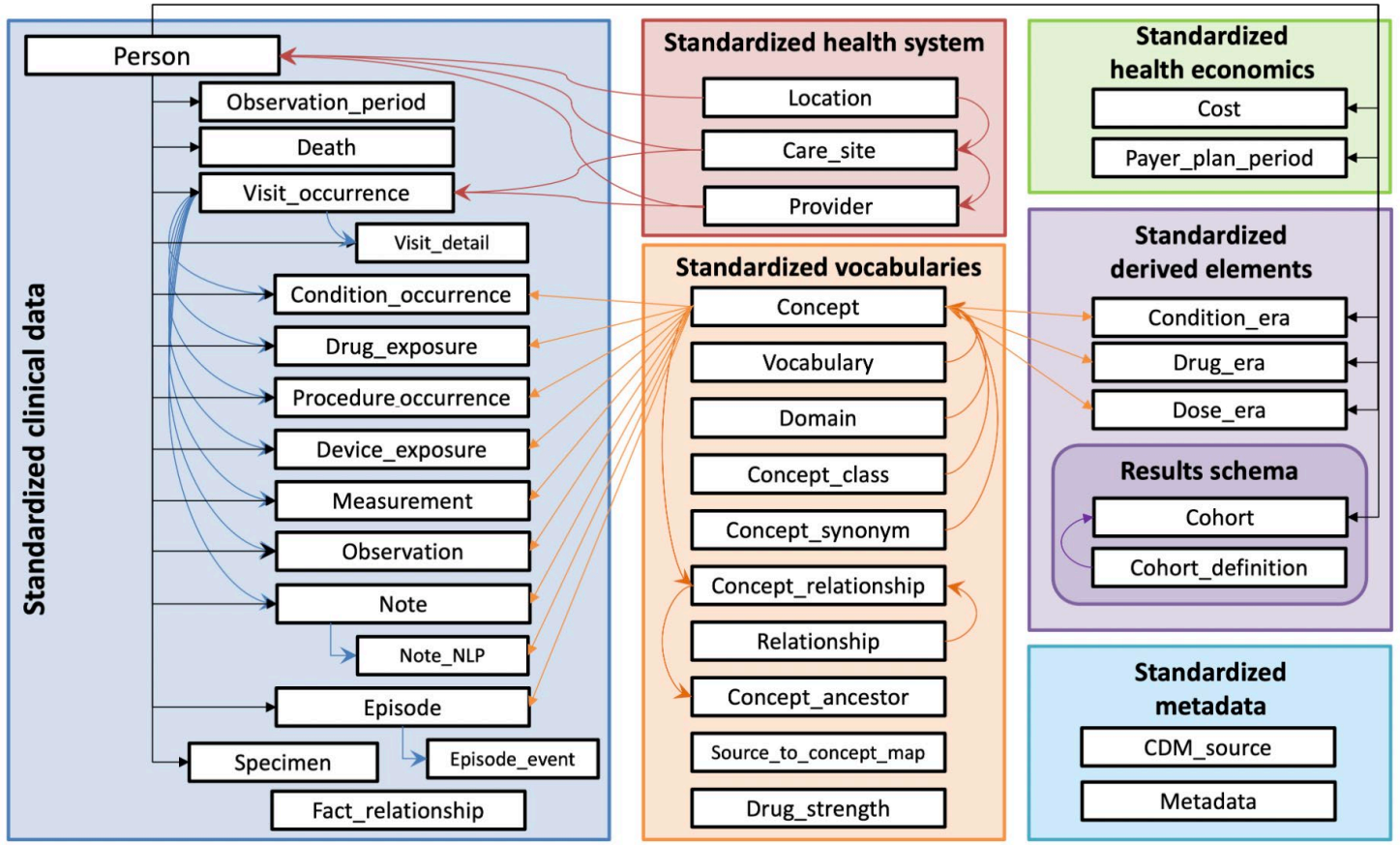
The Curated Data Repository (CDR) is a BigQuery data warehouse:

- 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

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

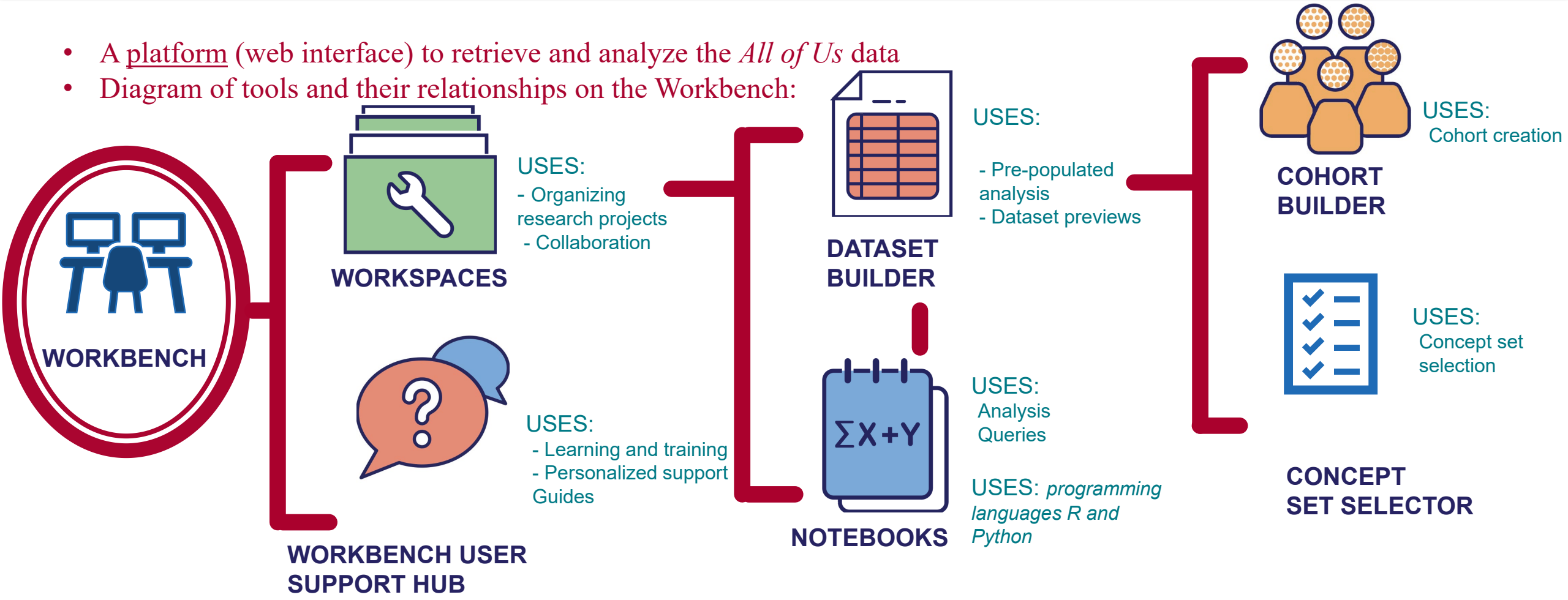
The *All of Us* Curated Data Repository has some data stored according to the **O**bservational **M**edical **O**utcomes **P**artnership (OMOP) **C**ommon **D**ata **M**odel (CDM) data standards from the **O**bservational **H**ealth **D**ata **S**ciences and **I**nformatics (OHDSI) vocabulary

(ohdsi.github.io/CommonDataModel/)



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:



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