Introduction to AIR·MS

AI Ready Mount Sinai

Patricia Kovatch
Eugenia Alleva
Herve Dibello
Andrew Deonarine
Jielin Yu

April 24, 2025

HPI·MS

Mount Sinai

Hasso Plattner Institute for Digital Health at Mount Sinai

Welcome and Introduction

Patricia Kovatch
Professor and Dean for Scientific Computing and Data

Overview

1. What is AIR·MS?
2. Data Modalities in AIR·MS
3. Requesting Access to AIR·MS
4. Using AIR·MS on the Minerva High Performance Computing (HPC)
5. About the AIR·MS Data
6. Documentation and Support
7. QA

https://labs.icahn.mssm.edu/airms/

What is AIR·MS?

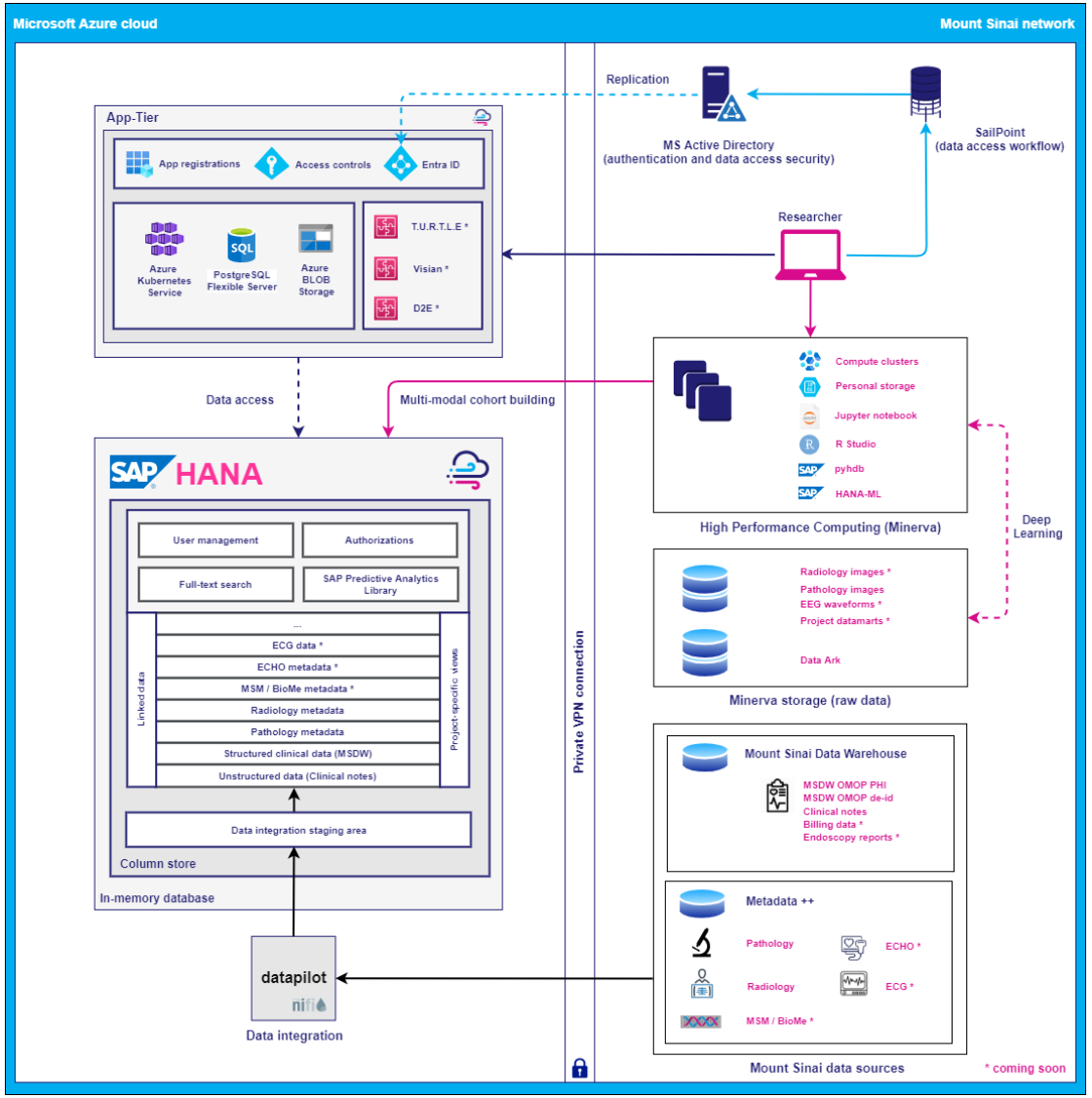
Artificial Intelligence-Ready Mount Sinai (AIR·MS) is a research platform that is composed of:

1) A (very fast) integrated database including Mount Sinai Data Warehouse (MSDW), Pathology and Radiology metadata; and the included data is growing

2) A Research Environment that allows interactions with the AIR·MS database from Python or R

3) An Application Tier to host a growing number of applications – including cohort building tools and annotation apps

What is AIR·MS?



Data Modalities in AIR·MS

Currently Available Modalities:

- Mount Sinai Data Warehouse (MSDW), both containing protected health information (PHI) and DeID (deidentified) Observational Medical Outcomes Partnership (OMOP)-mapped electronic health record (EHR)
- Pathology Metadata
- Radiology Metadata

Work in progress: BioMe/Sinai Million, electrocardiogram (EKG), Echocardiography, GI Research Database, electroencephalogram (EEG), Endoscopy & Colonoscopy Reports

All modalities are stored in separate database schemas, and access is granted to each schema individually based on Institutional Review Board (IRB)

Request Access to AIR·MS in Sailpoint

1) Obtain Institutional Review Board (IRB) approval for your project if you want to access PHI data, including this indication "*we will use the AIR·MS platform (IRB # 20-01288) to access and store our data*"

2) Request Minerva/High Performance Computing (HPC) account

3) Request access to specific modalities (i.e. schemas) on SailPoint

4) After access approval you can get started with the data using our Getting Started Guides:

```
git clone https://github.mountsinai.org/AIRMS/airms-researcher-tutorials-minerva.git
```

Use AIR·MS on Minerva HPC

High Performance Computing and Data Ecosystem - Minerva Supercomputer Cluster

AIR·MS

- Minerva Supercomputer provides scalable high performance computational and data infrastructure
- Use Minerva to perform large computing jobs and run large applications to advance and accelerate scientific discovery at Mount Sinai
- Compute nodes
  - >24,000 Intel cores
  - >350 NVIDIA graphic processing units (GPUs)
- Storage
  - 40 petabytes (PB) of usable parallel storage
  - TSM archiving for off-line storage
- Software
  - >1,000 applications

General Minerva Information

AIR·MS

- The Minerva website is: https://labs.icahn.mssm.edu/minervalab/
- Contains the documentation for the features and access to the forms

- Access to Minerva requires a Minerva userid
  - You are not automatically assigned one, you will need to apply for it
  - Username will be the same as your Mount Sinai Login id
  - Link to form on the Scientific Computing home page or https://tinyurl.com/ycx3r5ek

How to Run Jupyter Notebook on Minerva

Option 1 : On-the-fly Jupyter Notebook in a Minerva job

https://tinyurl.com/27erdpjp

Scripts in /usr/local/bin on head nodes only:

- `minerva-jupyter-module-web.sh`
  - Runs Jupyter installed in module python/3.12.5
  - Can access modules on Minerva
  - Can load users' conda environments (Jupyter must be installed in conda env)
    - `minerva-jupyter-module-web.sh -mm anaconda3/2024.06 -env ENV_NAME`
    - Install jupyter in your conda env: *`conda install jupyter`*
    - Do not have any conda env activated before running the script.
- Add --help to the script to get help message/usage

How to Run Jupyter Notebook on Minerva (Cont'd)

AIR·MS

Option 2 : Jupyter Notebook in a batch job

- Primarily, the nbconvert tool allows you to convert a Jupyter .ipynb notebook document file into another static format including HTML, LaTeX, PDF, Markdown, reStructuredText, and more. nbconvert can also add productivity to your workflow when used to execute notebooks programmatically.
  https://tinyurl.com/5batvyxu
  Use the commands below in your job script:

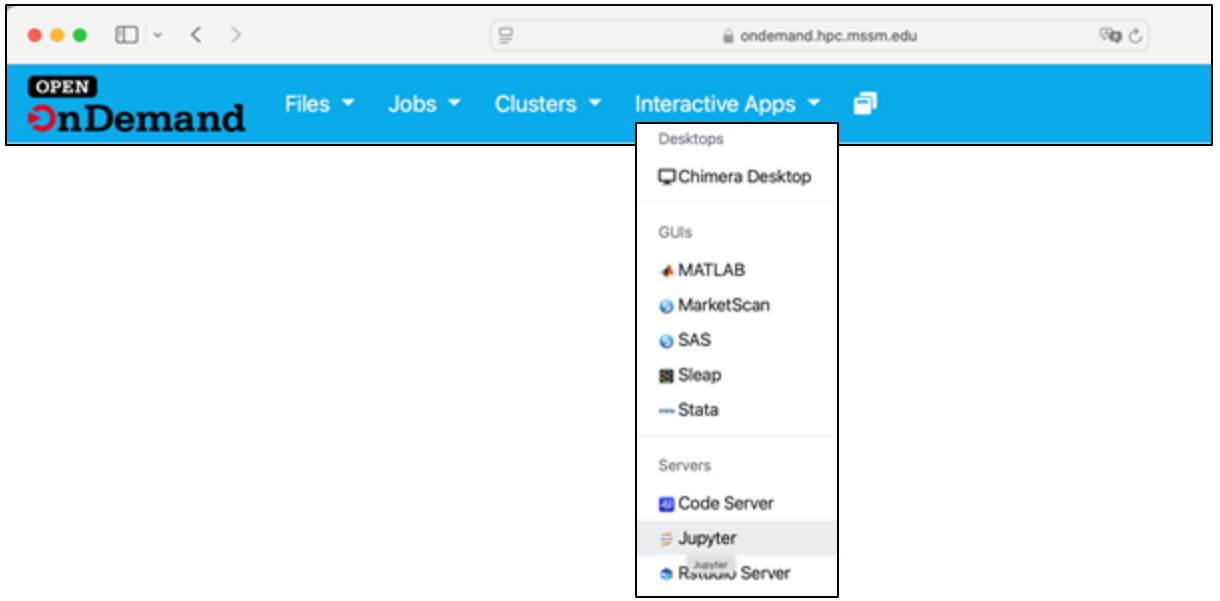
```
ml python/3.12.5
jupyter nbconvert --to notebook --execute myfile.ipynb
```

- The result is saved in a file named **myfile.nbconvert.ipynb**

How to Run Jupyter Notebook on Minerva (Cont'd)

Option 3 : Via Open OnDemand

- Point your browser to https://ondemand.hpc.mssm.edu/pun/sys/dashboard
- Login with your Mount Sinai username/password and multi-factor authentication device.
- In the User Interface Dashboard, Click 'Interactive Apps> Jupyter' on the menu bar



Connecting to AIR·MS

Connect to AIR·MS on Local Device

AIR•MS

Connecting to AIR.MS from Local Device

```
[ ]: !pip install airms-connect --index-url=https://airms_python_packages_test:2uqmjg6w6twinvsrzew3vaxzxvvquc2vgp64igvfzkseuxlgwreq@pkgs.dev.azure
```

```
[ ]: # import airms connection library
from airms_connect.connection import airms_connection
```

```
[ ]: # initialize connection
airms = airms_connection()
```

```
[ ]: # connect to AIR.MS
airms.connect(schema='CDMDEID')
```

```
[ ]: # query the database
person_df = airms.conn.sql('SELECT TOP 10 * FROM PERSON').collect()
```

Connect to AIR·MS on Minerva (Interactive Session)

AIR•MS

Connecting to AIR.MS from Minerva

```
[ ]: !pip install airms-connect --index-url=https://airms_python_packages_test:2uqmjg6w6twinvsrzew3vaxzxvvquc2vgp64igvfzkseuxlgwreq@pkgs.dev.azure
```

```
[ ]: # import airms connection library
from airms_connect.connection import airms_connection
```

```
[ ]: # initialize connection
airms = airms_connection()
```

```
[ ]: # establish an ssh tunnel with a Minerva login node
airms.on_minerva(login_host_name='li04e04')
```

```
[ ]: # connect to AIR.MS
airms.connect(schema='CDMDEID')
```

```
[ ]: # query the database
person_df = airms.conn.sql('SELECT TOP 10 * FROM PERSON').collect()
```

Connect to AIR·MS on Minerva

AIR•MS

Live Demo

Using the Data

Using the Data: Introduction to OMOP

AIR·MS

- Data is stored in “OMOP” format
- **OMOP = Observational Medical Outcomes Partnership**
- Work on OMOP originally started in 2008, through a coloration between the NIH, FDA, and pharmaceutical industry
- Emerged out of an effort to standardize data for pharmacoepidemiology
- The initial work was completed in 2013

Using the Data: Introduction to OMOP (Continued)

AIR·MS

- After the initial work completed in 2013, the OMOP group then became the Observational Health Data Sciences and Informatics (OHDSI) group (pronounced “Odessey”)
- The common data model (CDM) for OHDSI is OMOP
- Many organizations throughout the US, Europe, and Canada use OMOP
- OHDSI performs several activities with OMOP, including governance, development, improvements, and assessing utilization
  - They also maintain several tools that work with OMOP for data analysis
  - A significant project is the OMOP vocabulary
- In the US – OMOP leadership is located at Columbia
  - Visit https://ohdsi.org

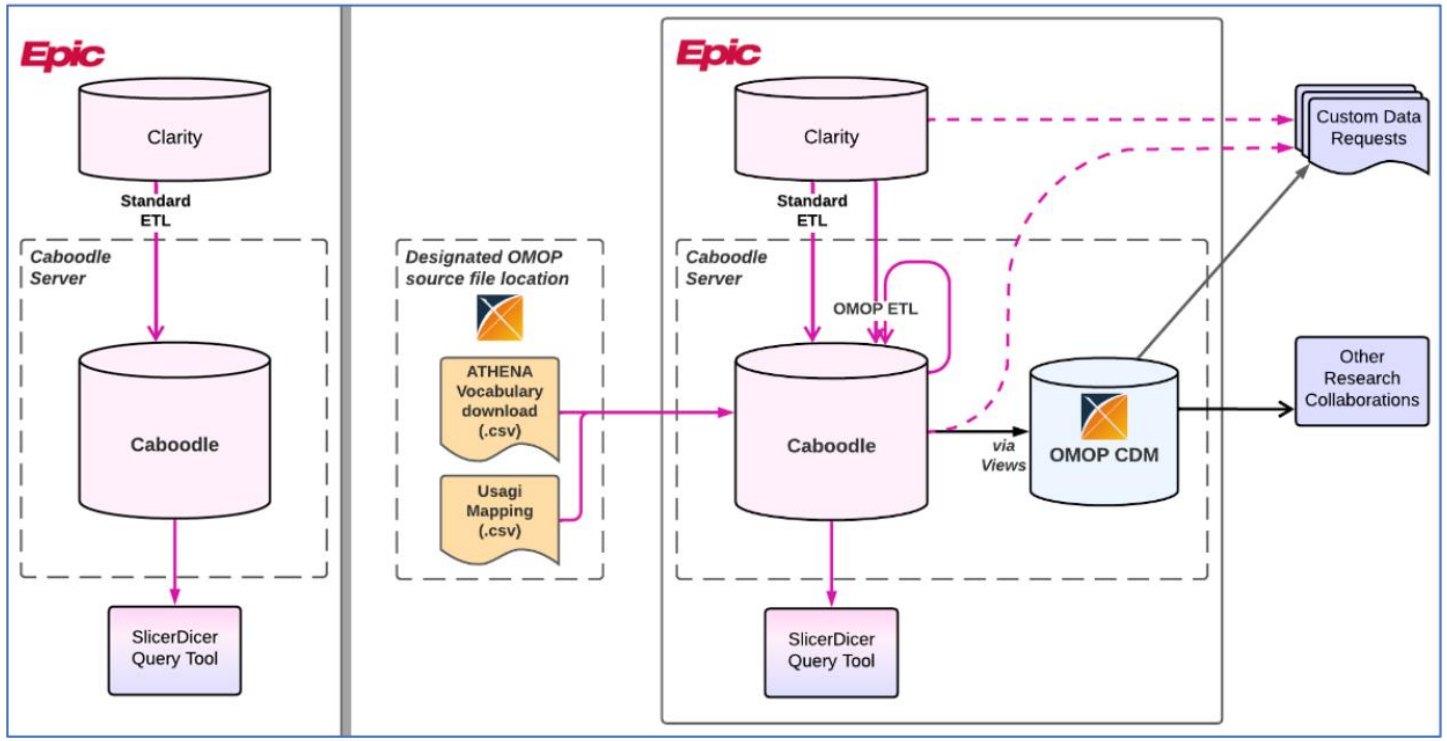
Using the Data: Who Uses OMOP?

AIR·MS

- Several large collaborations use OMOP:

- Federal observational data research usually involves OMOP formatted data

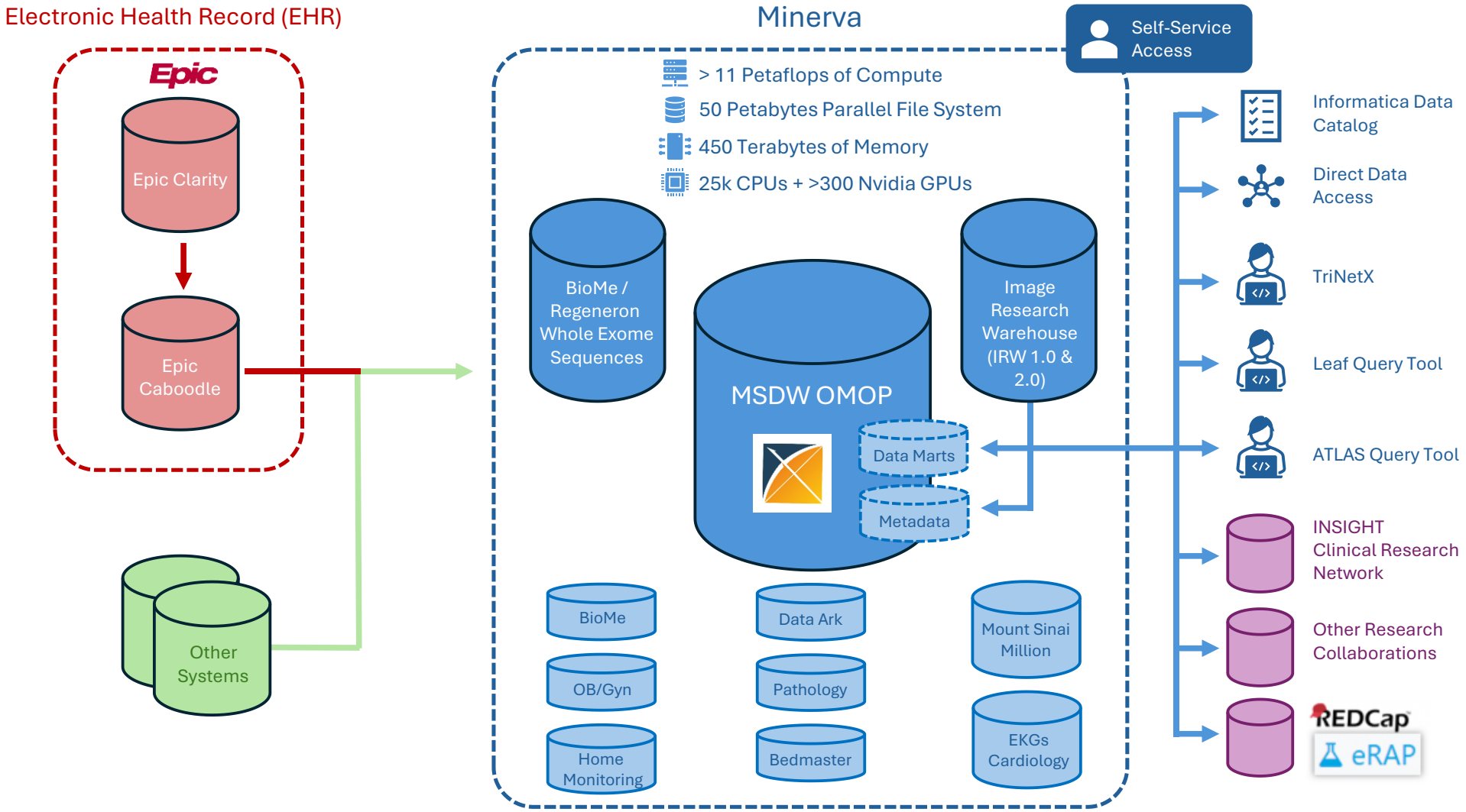
Using the Data: How Does OMOP Work?



https://www.ohdsi.org/wp-content/uploads/2023/10/10-Willett-BriefReport.pdf

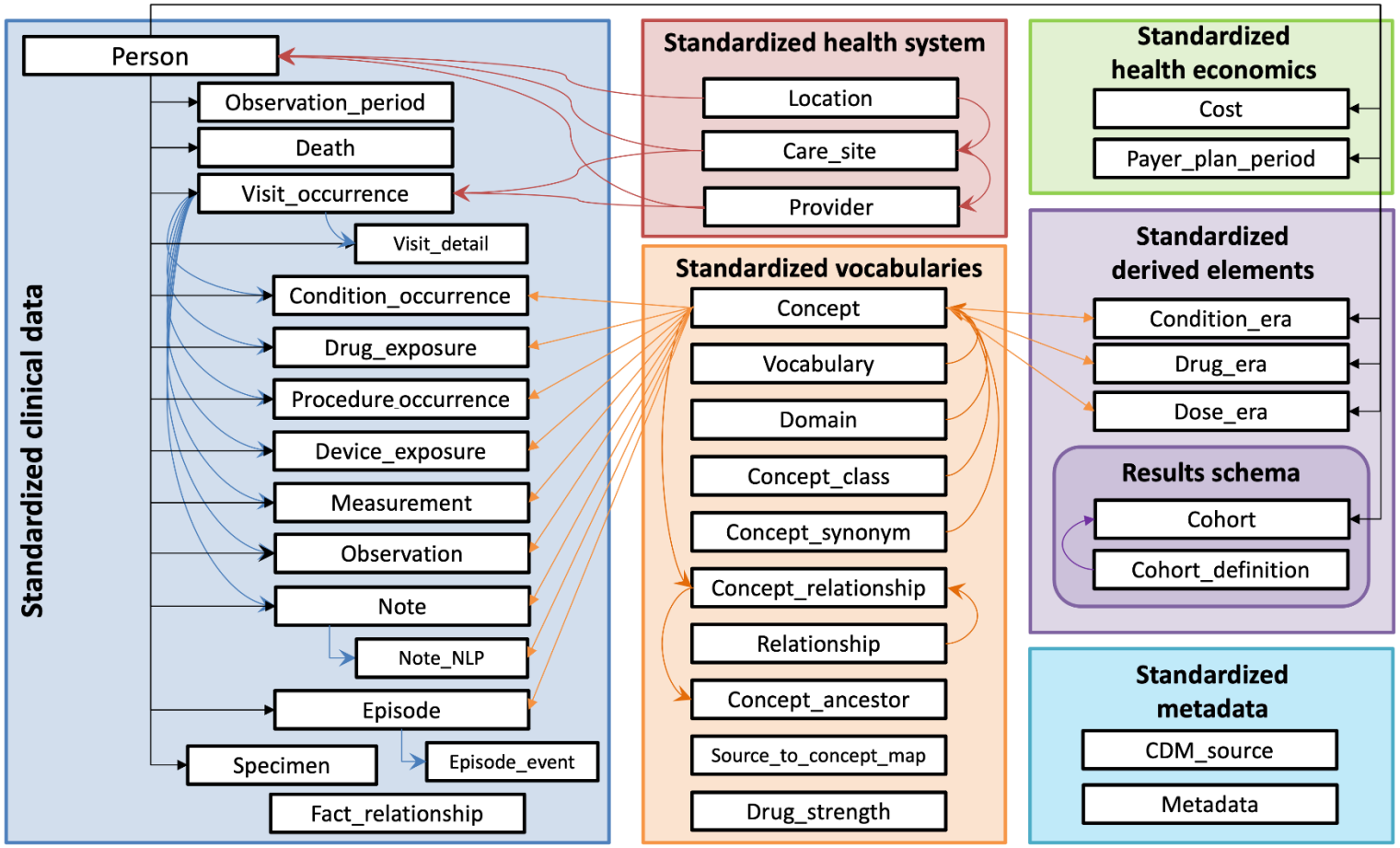
- Data is extracted from Caboodle into the OMOP Common Data Model (CDM)

Using the Data: How Does OMOP Work? (Cont'd)



Using the Data: How Does OMOP Work? (Cont'd)

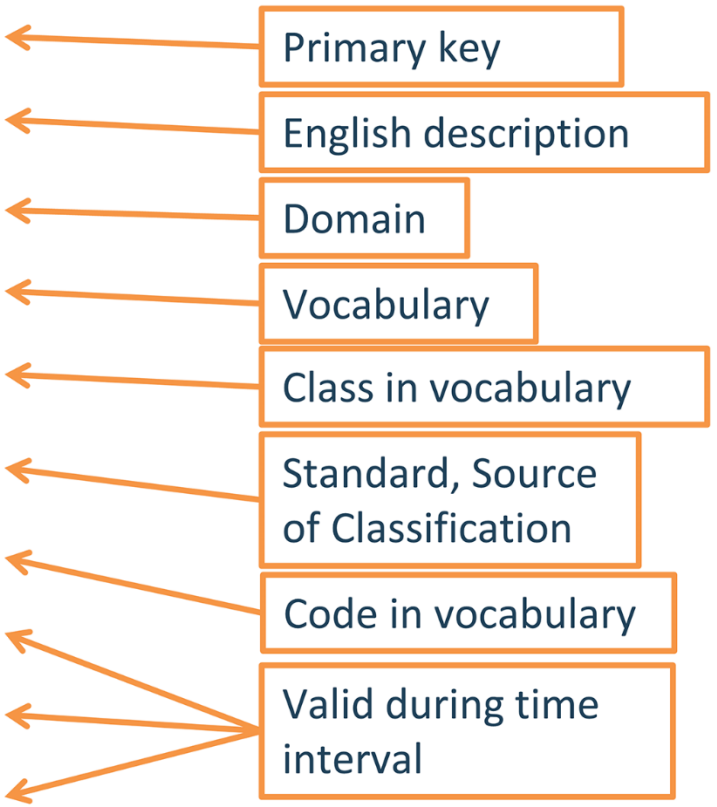
- The OMOP CDM consists of about 35 tables
- Very simplified compared to the ~20,000 tables in Clarity



Using the Data: Querying and Retrieving Data

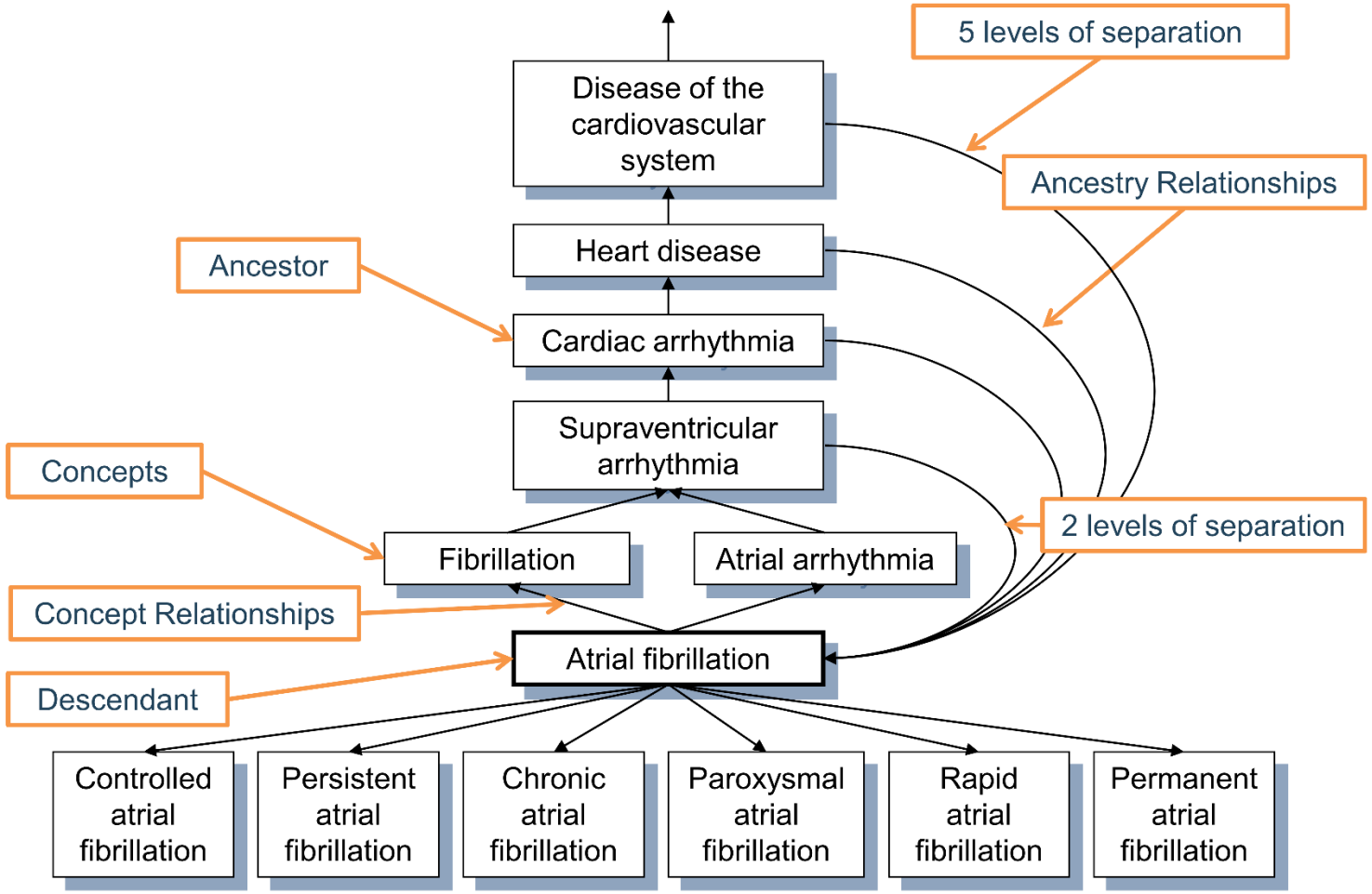
- Foundational to OMOP is a “Concept” (stored in the CONCEPT table)
- Concept domains: “Condition,” “Drug,” “Procedure,” “Visit,” “Device,” “Specimen,” etc.

CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	



Using the Data: Querying and Retrieving Data (Cont'd)

- Different terminologies are mapped to the Systematized Nomenclature of Medicine (SNOMED), and there is an ontology which defines concept relationships



Using the Data: Querying and Retrieving Data (Cont'd)

AIR·MS

- Here is a query for patients with new onset diabetes
- Use the table "condition_occurrence"
- This type of query will be very fast on AIR·MS compared to a standard database
- Excluded some codes with "…" for brevity

```
-- Cohort definition: Patients with new onset diabetes
SELECT DISTINCT c.person_id
FROM
-- Include all condition occurrences of diabetes
condition_occurrence c
WHERE
  -- Limit to diabetes diagnosis
  c.condition_concept_id IN (
                             -- Relevant clinical concepts for diabetes diagnosis:
                             31967, -- Diabetes mellitus type 1
                             36684827, -- Diabetes mellitus type 1 in remission
                             31968, -- Diabetes mellitus type 2
                             ...
                              -- Diabetes mellitus unspecified complication
                              )
  AND NOT EXISTS (
                 -- Exclude patients with any diabetes diagnosis prior to a 365-day lookback period
                 SELECT 1
                  FROM   condition_occurrence c2
                  WHERE  c2.person_id = c.person_id
                         AND c2.condition_concept_id IN ( 31967,
                                                          -- Diabetes mellitus type 1
                                                          36684827,
                                                          -- Diabetes mellitus type 1 in remission
                                                          31968,
                                                          -- Diabetes mellitus type 2
                                                          4304378,...
                                                        )
                         AND c2.condition_start_date < c.condition_start_date -
                                                       interval '365' day);
```

Using the Data: Querying and Retrieving Data (Cont'd)

- Major things to think about when querying data:

  - **1. Phenotypes** – when you're researching a disease (ex. Lyme disease) – just looking for a particular International Classification of Diseases (ICD) code or disease keyword ("lyme") is not enough.
    - You need to consult clinical practice guidelines
    - Sometimes there's an initial diagnosis that's wrong – you may need to look for lab results (ex. 2 separate diagnoses of diabetes within 6 months of each other, and an HbA1c level)
    - So, you might write a query that looks for patients with three Type 2 diabetes codes within a 6-month period, and two HbA1c levels above a certain threshold, rather than simply looking for patients with one Type 2 diabetes code
    - When possible, consult physicians and clinicians when devising how to identify patients with a particular disease or phenotype, peer-reviewed publications are also a good source of information
    - For example – NIH N3C specific codes for COVID-19: https://tinyurl.com/4dwruasx

Using the Data: Querying and Retrieving Data (Cont'd)

- Major things to think about when querying data:

  - **2. Drug prescription information** – pharmacoepidemiology can be a complex topic, just because someone was prescribed a drug does not mean they took it
    - If you are looking at drug interactions, just because someone is co-prescribed two drugs does not mean the potential interaction you saw was due to the drugs being taken at the same time
    - Often you will need to look at more complex patterns in information, including time-series models, to interrogate information

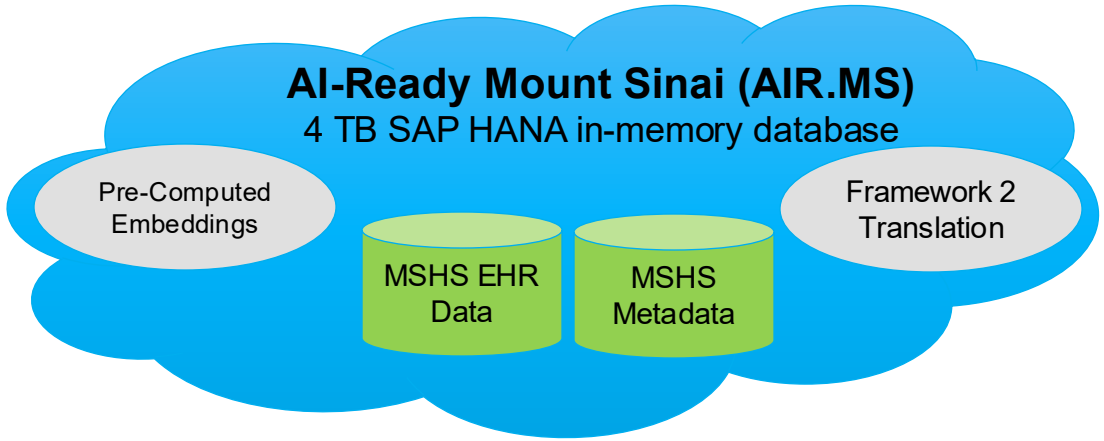
  - **3. Don't just keyword search in coding systems –** you will need to ensure that you have a list of synonyms for a clinical condition or concept
    - Searching International Classification of Diseases (ICD)-10 for "stroke" won't find "cerebral infarction"
    - Consult the literature or a clinician to determine what codes to include in your study

  - **4. You can use Large Language Models (LLMs) like ChatGPT to build queries, but be careful -** the code should be carefully checked, many LLMs are familiar with the OMOP format

AI-Enabling Infrastructure

Self-Service Access

- Data & Code Catalog
- Direct Database Access & Query
- Knowledge Graph
- Dashboards
- Direct Access & Open OnDemand
- Train, Test & Validate Models
- Store, Find & Share Models



Track Model Translation

Minerva Supercomputer
Confidential Compute
11 Petaflops of peak 64-bit compute
25k CPUs + >300 Nvidia GPUs
450 TB memory
50 PB parallel file system

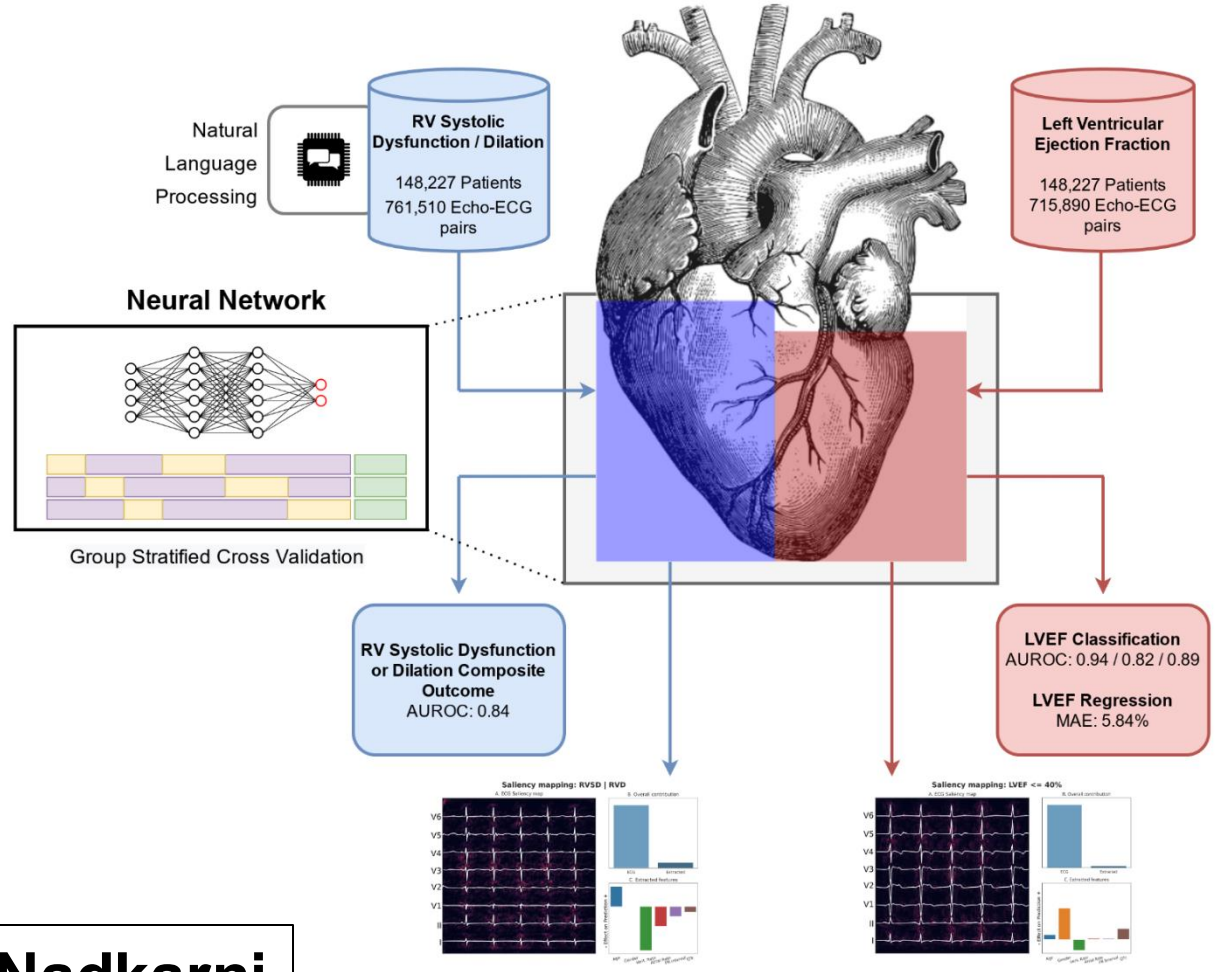
- Radiology & Pathology images
- EKGs
- Cardiology
- Home monitoring
- OB/GYN
- Bedmaster
- EHR
- MSDW

Regeneron Whole Exome Sequences & Somatic Return of Results

App Store
AI "As a Service" Tools
AI Build Environment

Utilizing AI on ECGs to Accurately Determine Heart Function

- Given a medical record number (MRN), the AI model gives a predictive probability of aortic stenosis, Chronic Obstructive Pulmonary Disease (COPD)
- AI model under Food and Drug Administration (FDA) review
- Training, inference performed on Minerva
- Validated the model by ensuring it performed the same in all patient subgroups (by age, sex, race)
- Being actively used in clinical care
- AIR.MS will be used to develop cohorts



PI: Girish Nadkarni

Documentation and Support

Documentation at https://tinyurl.com/v9hmdcyw

FAQ page at https://tinyurl.com/yc6a9j97

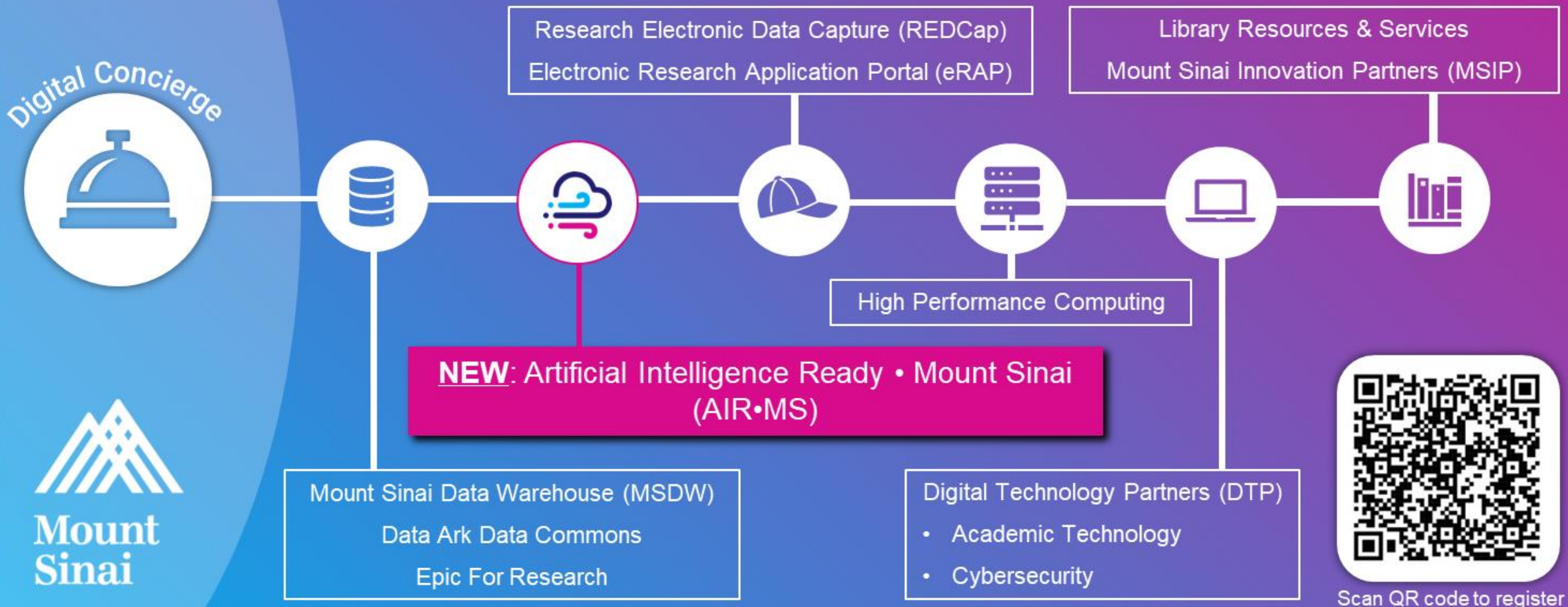
Support:

General questions and inquiries about AIR·MS	airms-info@mssm.edu
Issues using AIR·MS or during your onboarding process	airms-support@mssm.edu
Question about using Minerva HPC	hpchelp@hpc.mssm.edu

Online Digital Concierge

(Every Wednesday 3:30–4:30 pm)

Find out more about the following digital resources for research



Mount Sinai

Scan QR code to register

Review of Selected Training Offered in Spring 2025

Session	Topic
1	Minerva Intro
2	Load Sharing Facility (LSF) Job Scheduler
3	Introduction to GPU/AI resources on Minerva
4	Accelerating Biomedical Data Science with GPUs: Practical Approaches And Tools
5	Leveraging Large Language Models in Biomedical Research
6	Access Minerva via web browser Open OnDemand
7	How to Accelerate Genome Analysis Toolkit (GATK) by using Parabricks
8	Introduction to Data Ark
9	De-identified Digital Pathology Training Session

Spring Training sessions have ended but training resources from these sessions are available here: https://tinyurl.com/45uv98w9

Dates for Fall Training To Be Announced

Spring 2025 Town Halls

Coming soon: Spring 2025 Town Halls

Summarizing usage metrics, accomplishments, updates and roadmaps for each of the services we offer.
See details and registration links below:

- The **REDCap** Town Hall will be held **Friday, April 25th, Noon - 1 pm**. Click **here** to register for the Zoom webinar.
- The **eRAP** Town Hall will be held **Friday, May 2nd, Noon - 1 pm**. Click **here** to register for the Zoom webinar.
- The **Mount Sinai Data Warehouse (MSDW)** Town Hall will be held **Thursday, May 15th, Noon - 1 pm**. Click **here** to register for the Zoom webinar.
- The **Minerva HPC and Data Ark** Town Hall will be held **Friday, May 23rd, 1 pm- 2 pm**. Click **here** to register for the Zoom webinar.

Please visit our website for more information.

Q & A


Thank You
HPI·MS
Mount Sinai
Hasso Plattner Institute for Digital Health at Mount Sinai