Data Ark Data Commons Town Hall May 2023

Yiyuan Liu, Bioinformatician, Scientific Computing and Data
Jielin Yu, Computational Scientist, Scientific Computing and Data
Maria Julia Castro, Project Manager, GGS
Lili Gai, Director, High Performance Computing
Patricia Kovatch, Dean for Scientific Computing & Data & Professor, GGS

May 3rd, 2023

Mount Sinai

Outline for Today:

- Overview of Data Ark and Updates
- New Data Sets
- 2023 User Survey Results
- Data Ark Usage Summary
- 2023 Initiatives and Roadmap

Meet the Data Ark Operations Team

Patricia Kovatch

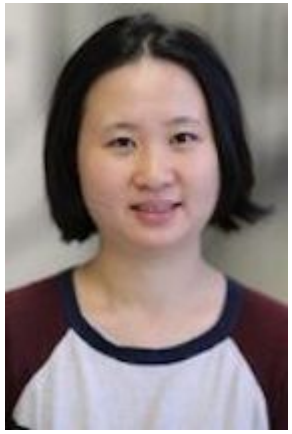
Professor and Dean for Scientific Computing and Data

Ranjini Kottaiyan, MBA

Senior Director

Maria Julia Castro, MS

Project Manager



Yiyuan Liu, PhD

Bioinformatician

Jielin Yu, PhD

Computational Scientist

Lili Gai, PhD

HPC Director

What is Data Ark?

Data Ark Mount Sinai Data Commons
Funded by **GGS** and **Scientific Computing**

Increasing the power, pace and relevance of our science

Power: Our researchers could be using ~20x more data

Pace: Users will have rapid access to huge powerful data

Relevance: Sinai can be a world leader in biomedical science

- What is the Data Ark?

–**Space on Minerva** to host all frequent-use research data sets (UK Biobank, GTEx, COVID Biobank…)

–**A team** of data scientists/engineers to manage resource, process data, simplify access process

–**An opportunity** for a step-change in the power and pace of Sinai research true 'big data science'

Data Ark is part of Mount Sinai's Computational and Data Ecosystem

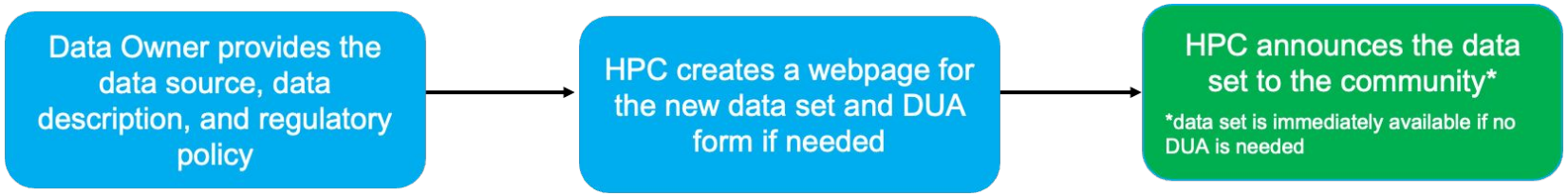
- Co-locating images, genomic, EHR and other data sets with the compute enables large-scale, multi-modal and multi-scale analyses

- Utilizing high-performance computing accelerates analyses

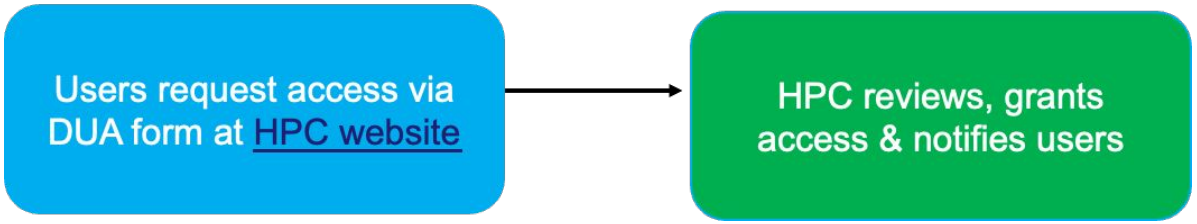
- Enabling researchers to directly query data maximizes accessibility

Standard Data Ark workflow: streamlined user access within 24 hours

Onboarding a Data Set



Granting User Access



Data Ark data commons
Streamlined Data Access within 24 hours

- **There are 18 data sets hosted under Data Ark currently**
  - Immediate access to 11 public-unrestricted data sets
  - Access within 24 hours to five Mount Sinai-generated data sets

Immediate Access

Public Data Sets

- 1,000 Genomes Project
- GTEx
- GWAS Summary Stats
- gnomAD
- The Cancer Genome Atlas (TCGA)
- eQTLGen
- UKBB-LD
- LDSCORE
- BLAST
- Reference Genome
- Genebass

Access within 24 hours

Mount Sinai Generated Data

- The CBIPM-BioMe Data Set
- MSDW COVID-19 EHR Data Set
- Mount Sinai COVID-19 Biobank
- The Living Brain Project
- STOP COVID NYC Cohort

Restricted Access

Public Data Sets

- UK Biobank

School-Acquired Data Sets

- MarketScan®

New Data Sets Updates

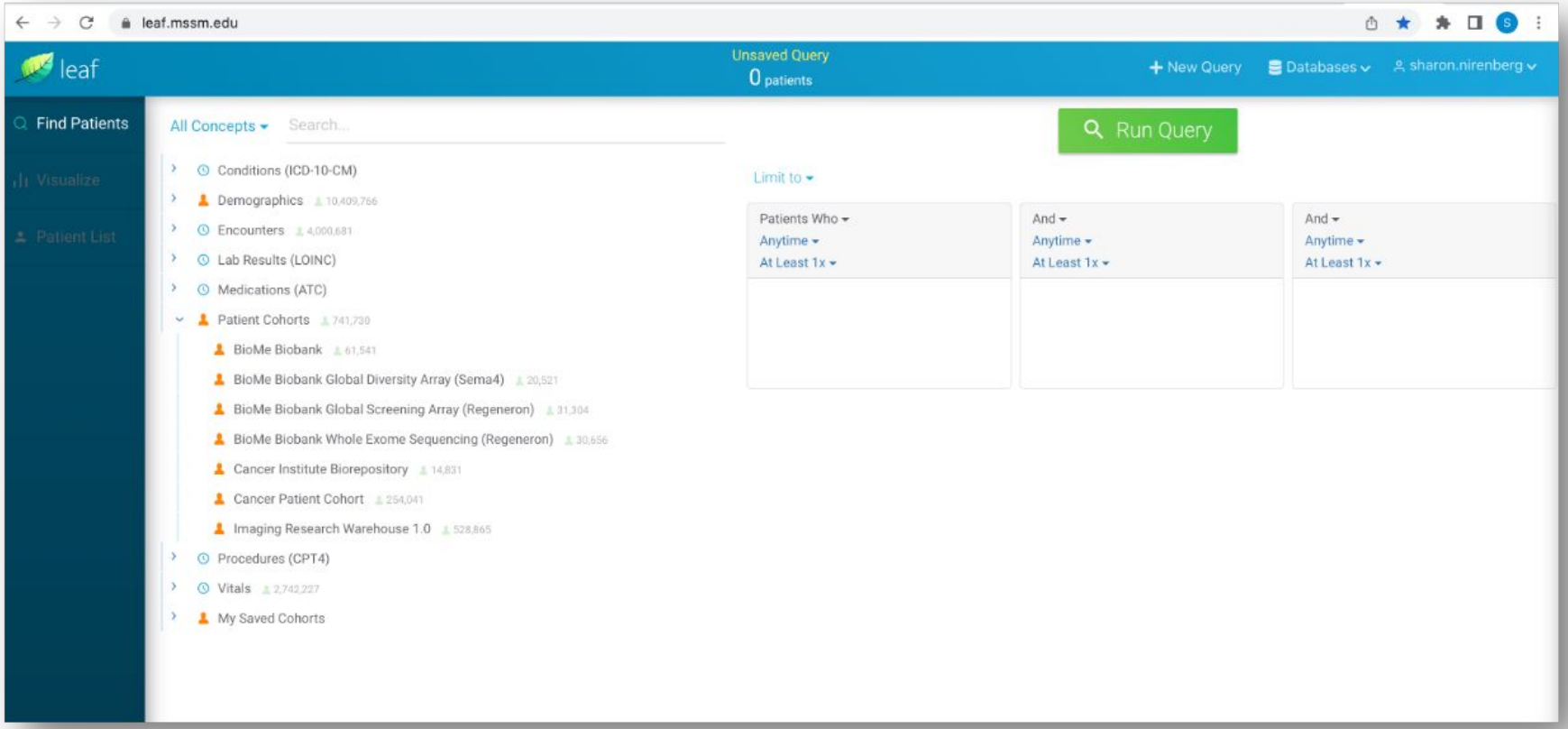
- **New Data Sets Added**
  - BioMe from CBIPM: De-identified phenotypic and genotype (microarray + WES data)
  - MarketScan
  - Genebass
  - UK biobank
    - Adding new applications per PI request
  - GWAS
    - Oxford Brain Imaging Genetics Server - BIG40

New Data Sets 1

- **CBIPM - BioMe Data Set**
  - The CBIPM-BioMe data set currently includes:
    - the BioMe Global Screening Array (GSA) (from Regeneron)
    - Global Diversity Array genotyping array (GDA) (from Sema4)
    - Whole Exome Sequencing Data (Regeneron)
    - BioMe Epic EHR Data Mart
  - **All the data are anonymized**

  - The CBIPM data set is at its freeze V2 version currently:
    - a combined set of 53,982 genotyping-array samples imputed with the 1000G- and TOPMed reference panel
    - V2 consists of additional 22,299 BioMe Biobank samples with 1.6 million typed variants based on the Illumina Global Diversity Array genotyping array (GDA) + the former release of 31,683 imputed samples based on the Illumina Global Screening Array (GSA) with ~650k typed variants

Institutional Patient Cohorts are Searchable in Leaf



Use Leaf to query the Cancer Patient, BioMe or IRW Cohorts

Leaf – Patient Cohorts

Patient Cohorts on Leaf	Description
BioMe Biobank	Individuals with blood and/or urine samples in Mount Sinai's BioMe Biobank
BioMe Biobank Global Diversity Array	Individuals with blood and/or urine samples in Mount Sinai's BioMe Biobank and DNA analyzed with Illumina's Global Diversity Array by Sema4
BioMe Biobank Global Screening Array	Individuals with blood and/or urine samples in Mount Sinai's BioMe Biobank and DNA analyzed with Illumina's Infinium Global Screening Array by Regeneron
BioMe Biobank Whole Exome Sequencing	Individuals with blood and/or urine samples in Mount Sinai's BioMe Biobank and Whole Exome Sequencing (WES) data generated by Regeneron
Cancer Institute Biorepository	Pathology tissue repository representing over 50 primary tumor sites
Cancer Patient Cohort	Individuals in the Mount Sinai Health System who have been diagnosed with cancer
Imaging Research Warehouse 1.0	Multi-modal radiology images collected in the Mount Sinai Health System from 2017 to 2021

New Data Sets 2

- **MarketScan® Research Databases**
  - MarketScan® Research Databases provides one of the longest-running and largest collections of proprietary de-identified claims data for privately and publicly insured people in the U.S.
  - MarketScan Commercial Claims and Encounters Database (CCAE) (2013 - 2021) – contains data from active employees, early retirees, COBRA continuees, and dependents insured by employer sponsored plans (i.e., individuals not eligible for Medicare).
  - MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR) (2013 - 2021) – created for Medicare-eligible retirees with employer-sponsored Medicare Supplemental plans. This database contains predominantly fee-for-service plan data.
  - The license will be expiring 11/04/2023
    - Will depreciate without additional funding in 11/04/2023
    - Will send the current user groups a survey on sharing the license cost
      - Cost depends on the years needed, e.g. Commercial and Medicare 2017-2021 is $45,156.

New Data Sets 3

- **Genebass**
  - Genebass is a resource of exome-based association statistics, made available to the public. The dataset encompasses 4,529 phenotypes with gene-based and single-variant testing across 394,841 individuals with exome sequence data from the UK Biobank
- **GWAS - Oxford Brain Imaging Genetics Server - BIG40**
  - This open data server contains results from GWAS of almost 4,000 imaging-derived phenotypes from the multimodal brain imaging in UK Biobank.
  - A major update to the original BIG server, using data from the 40,000 subject imaging data release from early 2020.
  - Discovery sample size was 22,138 and the replication sample 11,086. Chromosomes 1:22 and X are included, resulting in associations with 17,103,079 SNPs.

2023 User Survey Results

2023 Data Ark Survey Results and Questions

Questions:

1. How would you rate your satisfaction with the Data Ark: Data Commons data quality and availability at Mount Sinai?
2. How would you rate your satisfaction with Data Ark support?
3. What barriers exist preventing your usage of Data Ark?
4. Please share any additional comments

11 users responded (out of 55 Data Ark users)

- Developed a response and emailed HPC users as well as uploaded to Data Ark webpage

Thank you for your feedback!
This is the motivation for our 2023 Data Ark Roadmap!

We have all the responses to the comments on our website here.

Data Ark User Survey Results: Question 1

How would you rate your satisfaction with the Data Ark data commons data quality and availability at Mount Sinai?

	N%
Very Satisfied	36.3%
Satisfied	36.3%
Neither Satisfied Nor Dissatisfied	18.1%
Dissatisfied	0%
Very Dissatisfied	0.9%
Decline to Answer	0%

User satisfaction (>=Satisfied)
2022: **72.6%**

Data Ark User Survey Results: Question 2

How would you rate your satisfaction with Data Ark support?

	N%
Very Satisfied	36.3%
Satisfied	36.3%
Neither Satisfied Nor Dissatisfied	27.2%
Dissatisfied	0%
Very Dissatisfied	0%
Decline to Answer	0%

User satisfaction (>=Satisfied) 2022: **72.6%**

Data Ark User Survey Results: Question 3

What barriers prevent you from using Data Ark?

- Not as user friendly
- Lack of knowledge about the resource
- Would it be possible to post demo videos?

Response:

- We have put instructions on how to access each data set and how to use Minerva on Data-Ark website.
- There will be more promotions/marketing surrounding Data Ark. We are sending announcements for new data set to The Office of Research Services (ORS) and hpcusers email list. Additionally, we will hold workshops, training and seminars in the future to get users informed.
- We will hold training sessions with demo this fall and post recordings on Data Ark website.

Data Ark User Survey Results: Question 4

Additional feedback/comments?

- More support, more ways to filter the data
- Updating data from Open GWAS project
- Could Data Ark have a version and development log?

Response:

- Currently we don't have a tool supported for filtering the data. This is an excellent suggestion for future roadmap. We will discuss and try to implement something like data catalog with a searchable metadata index.
- We just added "The Oxford Brain Imaging Genetics Server – BIG40" as suggested and will keep adding more as needed. We will also update GWAS summary data to the upcoming new format as needed. And if you have a particular data, please let us know at data-ark-team@lists.mssm.edu.
- We have a page describing the data set and version on our website https://labs.icahn.mssm.edu/minervalab/resources/data-ark/. There are also README files inside the folder of each data set. We will consolidate those files to make it more user friendly.

Data Ark Usage Summary (Oct 2022 - Mar 2023)

Dataset	Size (GB)	# of unique users 2022	# unique users Q4 2022	# of unique users Q1 2023	# of times accessed
UK Biobank	11,900	32	17	9	40,403
1000 Genomes	143	28	14	2	8,387
GTEx	1,888	20	7	5	608
gnomAD	8,628	11	10	2	279
MarketScan® Research Databases	2,265	7	7	14	267
GWAS	6,826	17	8	2	204
Reference genome	143	7	7	4	104
UKBB-LD	2,867	8	8	2	96
COVID-19 Biobank	379	18	0	1	2
eQTLGen	39	8	1	0	2
STOP COVID NYC Cohort	1	16	0	0	0
MSDW COVID-19 EHR	2	16	0	0	0
TCGA	155	9	0	0	0
The Living Brain Project	1	1	0	0	0

Supported over tickets: 2022: **94;** Q1 2023: **17**

- Support also provided over emails.

Researcher Engagement Updates:

- Data Ark Slack Current Activity:
  - #general" channel now has 65 members!
  - Users are actively discussing possible additions/edits of data sets on the #general channel- thank you to the O'Reilly lab for being active on the channel!

- Presentations:
  - Data Ark Town Hall twice a year
  - Training session will occur twice a year starting this fall
  - The TCI Basic and Translational Committee presentation
  - Data Ark GGS department presentation in the fall

What's next

- Adding new data sets
  - MSDW de-identified OMOP dataset
  - De-identified digitized pathology slides
    - Links to slide metadata in MSDW
  - Bedmaster patient monitoring data
  - IRW 2.0 data available through Data Ark
  - Human Cell Atlas
  - Marketscan data set renewal
- Investigate/develop tools for more friendly data access such as indexing/catalogue
- Marketing and raise awareness
  - Offer training class every year and upload the slide/recording online
  - Revive HPC twitter account and post Data Ark related content
  - Collect publications that have been produced from Data Ark usage and post on website
  - Expand slack channel
  - Setting up grand rounds

How to interact with the Data Ark team

- Anyone can contact the Data Ark team with questions or ticket submissions by writing to data-ark-team@list.mssm.edu.

- Data Ark Slack community ---we have channels for every common data set. To join the channel, navigate to https://join.slack.com/t/data-ark/signup and sign up using your Mount Sinai credentials. You'll be able to start interacting with other researchers on common data sets right away.

- Click here for more information!

We want to hear from you!

Submit which data sets will be useful for your research:

Suggest a Data Set Survey

Acknowledgements

- Supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences, National Institutes of Health.

Thank you!

Data Ark Mission

The purpose of the Data Ark is to ensure that scientists at Mount Sinai have all the data they need to maximize the power of their science. These data should be easy-to-access, in analysis-ready form and be as large and diverse as possible.

Data Ark Data Sets

Launched March 2021!

The Data Ark is located on Minerva and the number, type, and diversity of data sets on the Data Ark will increase substantially in the coming months. The Data Ark consists of Restricted and Unrestricted data, data supplements, and cloud-hosted data. Core sets hosted on Minerva include:

Public Data Sets (unrestricted)

- 1,000 Genomes Project
- GTEx
- GWAS Summary Stats
- gnomAD
- The Cancer Genome Atlas (TCGA)
- eQTLGen
- UKBB-LD
- LDSCORE
- BLAST
- Reference Genome
- Genebass

Mount Sinai Generated Data (restricted)

- The CBIPM-BioMe Data Set
- MSDW COVID-19 EHR Data Set
- Mount Sinai COVID-19 Biobank
- The Living Brain Project
- STOP COVID NYC Cohort

Public Data Sets (restricted)

- UK Biobank

School-Acquired Data Sets (restricted)

- MarketScan®

Data Set Supplements – Minerva-hosted, Open Soon Through Data Ark (restricted)

- The Imaging Research Warehouse (IRW) 1.0
- CIB (Cancer Institute Biorepository)

Data Ark Org Chart