Minerva HPC and Data Ark Town Hall

Lili Gai, PhD, Director for High Performance Computing and Data
The Minerva HPC Team

May 23, 2025

Outline

1. Accomplishments and Updates
2. 2025 Roadmap
3. Appendix I – Minerva/Data Ark Usage
   Appendix II – 2024 Minerva User Survey Results
   Appendix III – 2024 Data Ark User Survey Results

Accomplishments and Updates

Accomplishments and Updates (Oct. 2024 - March 2025)

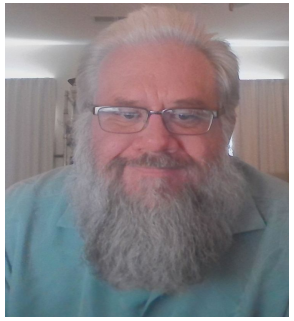
- Staff Updates
- $2M AI Mount Sinai (AIMS) NIH S10 Proposal to Add More Graphics Processing Units (GPUs)
- Updates on Developing National Institute of Standards and Technology (NIST) Special Publication (SP) 800-171 Environment on Minerva
- Launched New Annual Form to Attest to "Mount Sinai Health System AI Implementation and Use Policy"
- Updates: LSF Operation on Resource Allocation Limit
- Minerva Arion Upgrade to General Parallel File System (GPFS) Version 5.1.9
- Load Sharing Facility (LSF) Job Scheduler Patches to Fix Bugs
- Service Upgraded and Migrated to New Servers with Rocky OS 9.5 with New Features
- Updates on Root Cause for Unexpected Outage on Network File System `/home`
- Minerva Preventive Maintenance (PM) and Outage since Last TownHall
- Reminder on Minerva Data Backup Policy
- Training Offered Spring 2025
- Data Ark Updates

Welcome Our New High Performance Computing (HPC) Experts

Sumit Saluja, Senior HPC Administrator Joined Jan 2025

- Previously worked at Columbia University as Senior HPC Admin.
- 15 years of experience in HPC at Princeton University, NIH, and Columbia University, with expertise at designing, optimizing, and securing HPC clusters and big data storage systems with CI/CD automation.

Daniel Ross, Associate Director of HPC Joined Feb 2025



- Came from Oak Ridge National Laboratory, where he managed 28 HPC clusters while supporting both physics and medical research as Senior Lead.
- 20 years of experience in HPC at Oak Ridge National Laboratories, Emory University Rollins School of Public Health and Georgia Tech. with extensive experience with all standard procedures, security and compliance in HPC.

$2M AI Mount Sinai (AIMS) NIH S10 Proposal to Add More GPUs

Thank you for sending us your science story and your publications!!!

Thank You!

Status: Awarded on May16th 2025 !!!

Plan and Timeline:

1) Work with Vendor to get the node purchased and delivered to our Hess DC: 8~10 weeks
2) Work with Facility to get the power and water ready for the new rack: TBD
3) Installation: ~4 weeks

Background:

- We requested a new high-performance instrument with 6x DGX providing state-of-the-art GPU capability and capacity for AI Mount Sinai
- DGX B200 can deliver 3X the training performance and 15X the inference performance of DGX H100 (eg. B200 offer new capabilities such as floating point 4 bits (FP4)).
- *Impact Score: 11!!! with almost none weakness noted from the summary statement.*

$2M NIH S10 Proposal to Add More GPUs

- **The final AIMS machine will consist of 6x DGX nodes if awarded with a total of**
  - 48 NVIDIA B200 GPUs connected via NVLink and Infiniband NDR400
  - 672 Intel Xeon Platinum 8570
  - 9 TB of memory available on B200, and an additional 12TB available on the server nodes
  - Local high-speed NVME storage and DDR5 RAM enables caching of intermediate results

AIMS total system configuration	
Performance (petaFLOPS) [NVIDIA]	1.92 FP64 3.84 FP32 216 FP16 432 FP8 training 864 FP4 inference
# of GPUs	48 NVIDIA B200
GPU memory size (TB) and type	9 HBM3e
GPU memory bandwidth (TB/s)	384
NVLink bandwidth (TB/s)	87
# of CPU cores	672
System memory (TB)	12
# of nodes & type	6 Lenovo ThinkSystem SR780a V3 8-way GPU

Background: New Guidance on Securing Data from NIH Controlled-Access Genomic Data Repositories

On **January 25, 2025**, users who submit a new request or renew an existing request will be expected to secure data according to the updated NIH Security Best Practices:

- Approved Users will attest to NIH that the system storing human genomic data is compliant with National Institute of Standards and Technology (NIST) Special Publication (SP) 800-171.
- The attestation may vary. The Principal Investigator's (PI's) and institution's ability to attest is informed by a self-assessment of NIST SP 800-171.
- There are ~70 PIs who are using NIH-controlled data sets:
  - Most projects are using Minerva already.

Current Status: Meeting NIST SP 800-171 Requirements on Minerva

- Scientific Computing hosted *two Town Hall sessions* on NIH Genomic Data Repository in Mid Feb. providing status and plans.
- Developing a plan to implement NIST 800-171 on Minerva with several stages:
  - We have developed a System Security Plan (SSP) and a plan of action and milestones (POAM) for addressing controls that are not yet met. Digital and Technology Partners (DTP) approved those in 05/16/2025.
  - Working with the Grants and Contracts Office (GCO) and Legal to resolve to sign the new attestation for data request.
- When completed, **Minerva will be the only approved environment for NIH genomic data sets at Mount Sinai.**

Challenge to Meet NIST SP 800-171

- Implementing National Institute of Standards and Technology (NIST) Special Publication (SP) 800-171 cybersecurity requirements will take significant investment and time:
  - Security requirements for NIST 800-171 Revision 3 released in May 2024, introduced 17 control families and 97 controls including 509 assessment objectives.
  - Due to the NIST 800-171 requirements for all data encryption at rest and in flight, we need to dedicate storage and compute nodes with extra protection. This means slower compute and file system performance.
  - Scope and Cost
- We will leverage existing work done by other organizations to move forward at pace.
- A shared institutional responsibility requires joint efforts from Scientific Computing and Data, DTP, Grants and Contracts Office (GCO), and researchers.

Ongoing Work

Scientific Computing & Data

1. **Assess** Minerva on the NIST 800-171 for all 110 controls.
2. Develop a **System Security Plan (SSP)** to meet NIST 800-171.
3. **Create a Plan of Action and Milestones (POAM)** for implementation.
4. **Implement** new cybersecurity framework.
5. **Update** policies & procedures and website.
6. **Enact** and record new **Data Use Agreements (DUAs)**.
7. **Audit and maintain** infrastructure.
8. **Train** personnel and researchers.

DTP

1. **Provide** input on the DTP-managed resources for the NIST 800-171 SSP for Minerva (eg. Active Directory, user authentication, computer room security, supply chain, etc.)
2. **Validate** the SSP and POAM for Minerva.
3. **Audit and maintain** supporting infrastructure.
4. **Train** researchers.
5. **Update policies & procedures** (eg. Updated DUAs) and website.

GCO

1. **Track and manage** requests to access NIH-controlled genomic data repositories.
2. **Create** an attestation template for researchers, DTP and Scientific Computing (mentioning use of compliant IT environment, training, physical security, auditing, etc).
3. **Create** an updated DUA.

Researchers

1. **Complete** an attestation that their work meets NIST requirements.
2. **Store** the data on Minerva.
3. **Complete** new DUA.
4. **Complete** training to meet NIST requirements.

Support AI-Related Researches: Other Activities

Digital Concierge Service

- Added Artificial Intelligence (AI) – The AI team guides you to leverage AI tools for your research.
- Attendance: 41

Track and Understand the AI-related Applications and Research to Better Support Users

Monitor GPU Usage

- Reach out to top 5 graphics processing unit (GPU) Users on their use of GPUs

Launched New Annual Form to Attest to Institutional "MSHS AI Implementation and Use Policy"

Why?

- According to the institution's guideline and policy below, the ***DeepSeek model*** is not allowed to be used regardless of the sensitivity of the information.
- Prior to the use of any Artificial Intelligence (AI) Tool, you must present the AI Tool to the relevant committee (*e.g.*, the AI Review Board) for evaluation and approval with more details from Mount Sinai AI Implementation and Use policy: https://mshs.policytech.com/dotNet/documents/?docid=57430&app=pt&source=browse

Action Required

- As a result, DeepSeek model is prohibited on Minerva High Performance Computing (HPC) system and all other Sinai systems including your laptop. ***Please remove all related models and stop using it from Minerva.***
- Digital and Technology Partners (DTP) Cyber Security requires all Minerva users to attest to the "Mount Sinai Health System (MSHS) AI Implementation and Use Policy" annually. Please review and submit the agreement at https://forms.hpc.mssm.edu **by May 13, 2025**.
  - Campus network needed, or sign in through school VPN if off-campus
  - After signing in, select "MSHS AI Implementation and Use Policy".
- Any user who has not signed the form by then will have ***their Minerva accounts locked*** until the form is signed.

Updates: LSF Operation on Resource Allocation Limit

Goal:

- Stop users (needing heavy resource & misusing resource) from blocking the shared cluster resource and make better balance between users.
- Maintain high throughput for Minerva users.

Actions:

- We set dynamic global limits (keep monitoring and tune as needed):

  ✔ Max number of GPU running job slots per user is increased to 70 per user during the weekday (7AM -8PM), and 100 for the rest (actively changing from 20 ~ 100).

  ✔ Max number of running job slots is 4,000 per user during the weekday (7AM -8PM), and 8,000 for the rest (ranging from 2,000-8,000).

  ✔ Max requested memory of all jobs is 20TB - 40TB per user.
- We are actively monitoring every day with Grafana and cron jobs including the cluster load, usage per user, and wasted resource per user.
- We are working with and emailing users with a hybrid method: automatic emails and customized suggestion use Minerva appropriately.

Minerva Arion Upgrade to General Parallel File System (GPFs) Version 5.1.9

Scheduled Full Maintenance on Minerva on Thursday, Feb 20, from 8:00 AM to 8:00 PM.

Why & What are the Benefits?

- This is necessary to apply security patches, improve stability and benefit from the better features and support of the new Spectrum Scale version.
- We also need to reboot the enclosure with extended uptime to clear the sluggish enclosure services manager (ESM) response which can cause timeout and deadlock in accessing drives. This prevents us from performing online rolling upgrade.

What is the Plan?

- During this maintenance we will perform Spectrum Scale upgrade on the GPFs `/sc/arion` including the servers, enclosures and hard drives, and reboot the enclosure.

Load Sharing Facility (LSF) Job Scheduler Patches 2024 - 2025

We Made Two Quick Patches (~3 Hours in Total) in Dec. 2024 and Feb. 2025 to Solve:

- We noticed some users' LSF jobs exited after being launched with no reason recently. During this maintenance, we will apply patch LSF 10.1 Fix 602128 to address this issue. All the LSF daemon will be restarted on all the compute nodes.
- There is memory leaking from LSF daemons detected on the LSF master node. This will make the master server run out of memory quickly. Thus, we need to apply a quick patch recommended by IBM to fix the memory leaking issue.

How May this Impact Your Work?

1. During the installation of the patch, you might not be able to interact with LSF for submitting new jobs or query job/queue/host status such as bsub, bjobs, bqueues and bhosts etc.
2. All pending and running jobs are not affected.
3. All other Minerva service is not affected.

Root Cause for Unexpected Outage on Network File System (NFS) /home

- On **12/10/2024**, there was an unexpected 48-hour unprecedented NFS storage failure caused interruption to your research work.

- **Issues:** We hit a rare hardware failure which caused the home /hpc file system to be unavailable. An unexpected power module failed which caused both battery-driven storage controllers reboot. The controller has dual power supplies.
  - The failure of one power supply should not have crashed the storage controller.

- **Root Cause Analysis (RCA) from Vendor:** RCA for the volume failure has been isolated to a condition in which a power supply unit (PSU) failure can trigger ITN (Internal Temperature Notification) disconnects with drives. This type of failure is extremely rare with only 1 reported incident in the field to date for this product line.
  - The fix is a change in controller firewall (FW) to ensure that should a PSU fail in this fashion, the FW will isolate and fail the PSU thus preventing the ITN disconnects. (ESD-5054 – in progress) ETA- in progress.

Service Upgraded and Migrated to New Servers with Rocky OS 9.5

Globus for Data Transfer

- The globus is updated to new version 54-5.4.82.
- It is highly recommended to transfer files via the file manager with gloubs/GridFTP for better performance and fault tolerance. Currently, the file upload via HTTPS is limited to files < 16 GB.
- *UDT (UDP-based data transfer) is enabled in globus to help on transferring the large number of tiny files*.
- Aggregated Globus speed is up to 10Gb/s depending on the traffic load and network config at the other end

Shared Web Servers for User Websites

- *Web01 is migrated to two VMs,* each of which host the internal and external facing websites respectively for security, and improved reliability/ availability for these sites
- PHP/Apache upgraded to version Apache-2.4.62 and PHP-8.3.20

Cryosparc Server

- Migrated CryoSPARC to dedicated GPU nodes with Rocky 9.5
- Provided a guide as attached on how to manage CryoSPARC instance by your own group, so *users should be able to handle some simple issues of CryoSPARC* like restart or upgrade the instance

Cluster-Wide PM and Unplanned Outage in last six months

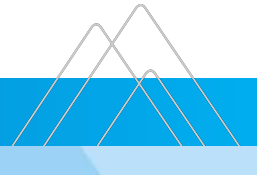
Full Minerva PM:

- **[1/23/2025]** 2.5 hours to perform a firmware upgrade on the drives for NFS storage /hpc
- **[1/20/2025]** 7.5 hours for Spectrum Scale upgrade on the GPFs file system

Some short windows on specific servers such as LSF patches, globus, web server and TSM

Unexpected system hang:

- **[12/10/2024]** Unexpected outage on Minerva home dir: 48 hours: 4 email updates
- **[1/23/2025]** GPFs outage for 1.5 hours: 2 email updates
- **[04/2025-05/2025]** Login nodes reboot due to high load and hardware issues



Reminder on Minerva Data Backup Policy

We are sending emails to reminder users about data backup policy on Minerva

This is a routine reminder on Minerva data backup policy.

1. **We do not backup any user files on any nodes including the private nodes. Please archive/backup your important files by yourselves.**
   a. We have included this in the message of the day (MOTD) after you log into Minerva, User Responsibilities and Acceptable Usage Policy, and Annual HIPAA Policy.
   b. We will keep reminding you every quarter via email.
   c. Please archive/backup your files following the guide at https://labs.icahn.mssm.edu/minervalab/documentation/access-tsm-with-command-line/

2. **Please don't set the permission of your Minerva files as rwx (read, write and execute) for everyone/others.**
   a. This can result in file deletion by others. Please double check your file permission on Minerva especially for your project directory.
   b. Limiting file permissions is the user's responsibility according to the annual HIPAA compliance requirement/training.

Minerva Training Offered Spring 2025

- Nine training sessions with training material (including slides & recording) online
- Online Digital Concierge every Weds (Data Ark, AI, HPC) from 3:30pm-4:30pm with 130 attendees

Session	Topic	Spring 2025 Date	Spring 2025 # of Attendees
1	Minerva Intro	Feb 13	71
2	Load Sharing Facility (LSF) Job Scheduler	Feb 20	47
3	Introduction to GPU/AI resources on Minerva	Feb 27	50
4	Accelerating Biomedical Data Science with GPUs: Practical Approaches And Tools	Mar 6	35
5	Leveraging Large Language Models in Biomedical Research	Mar 13	84
6	Access Minerva via web browser Open OnDemand	Mar 20	30
7	How to Accelerate Genome Analysis Toolkit (GATK) by using Parabricks	Mar 25	24
8	Introduction to Data Ark	Mar 27	23
9	Introduction to AIR.MS	April 24	129

Data Ark Updates

1. In Dec 2024, UK Biobank data offboarded from Data Ark due to data access now being restricted exclusively to its cloud-based platform.

2. In Dec 2024, Digital Pathology Slides accessible from Data Ark, while patients' EHR (electronic health records) data can be requested through MSDW (Mount Sinai Data Warehouse).

3. In Jan 2025, Data Ark IRB (Institutional Review Board) was renewed.

4. In Apr 2025, MarketScan data was migrated to the Research Academic IT team of DTP (Digital Technology Partners) due to security and compliance requirements.

2025 Roadmap

2025 Roadmap

- **Build NIST 800-171 Compliant System on Minerva According to the POAM.**
  - Enable Data encryption at rest: Encrypt File System with SED (self-encrypted disk) as CUI-Arion.
  - Enable Data Encryption in transit.
  - Enable Audit/logging with Spectrum Scale (GPFs) Data Management Edition.
  - Set up sustainable log server.
- **Jira Service Desk to replace User Osticket to enable status tracking and history tickets etc.**
- **Update Minerva documentation and FAQ on Website.**
- **Keep improving LSF config and monitoring.**
- **Explore confidential computing on Minerva Env.**
- **Explore how to help implement institutional AI-related policy.**
  - Host approved models in central repo.
- **Improve Minerva Fault Tolerance on NFS home.**
  - Install RedHat OS for our NFS servers.
  - Patch NFS with new fix on manual failed controllers.
- **Consolidate all VMs and renew VMware and hardware for all mgmt servers.**
- **Migrate Minerva authentication from Symantec to Azure MFA - Q4 2025.**
- **IRW (Imaging Research Warehouse) data: copying radiology images to Data Ark**
- **Launch Digital Pathology Slide Viewer**

Minerva Team Members

Patricia Kovatch
Professor and Dean for Scientific Computing and Data Div

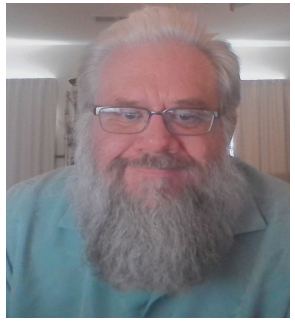
Lili Gai, PhD
High Performance Computing Director

Hyung Min Cho, PhD
Senior Computational Scientist

Jielin Yu, PhD
Senior Computational Scientist

Yiyuan Liu, PhD
Data Ark Bioinformatician

Shamimul Hasan, PhD
Senior AI/GPU Computational Scientist



Daniel Ross
Associate Director of HPC

Wei Guo, PhD
HPC Architect

Tejas Rao
HPC Architect

Eric Rosenberg
Senior System Administrator

Sumit Saluja
Senior HPC Administrator

Acknowledgements

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

Thank You

Icahn School of Medicine at Mount Sinai

Appendix I:
Minerva Usage
(Oct. 2024 – March. 2025)

Minerva Usage Summary (Oct.1 2024- Mar. 31, 2025)

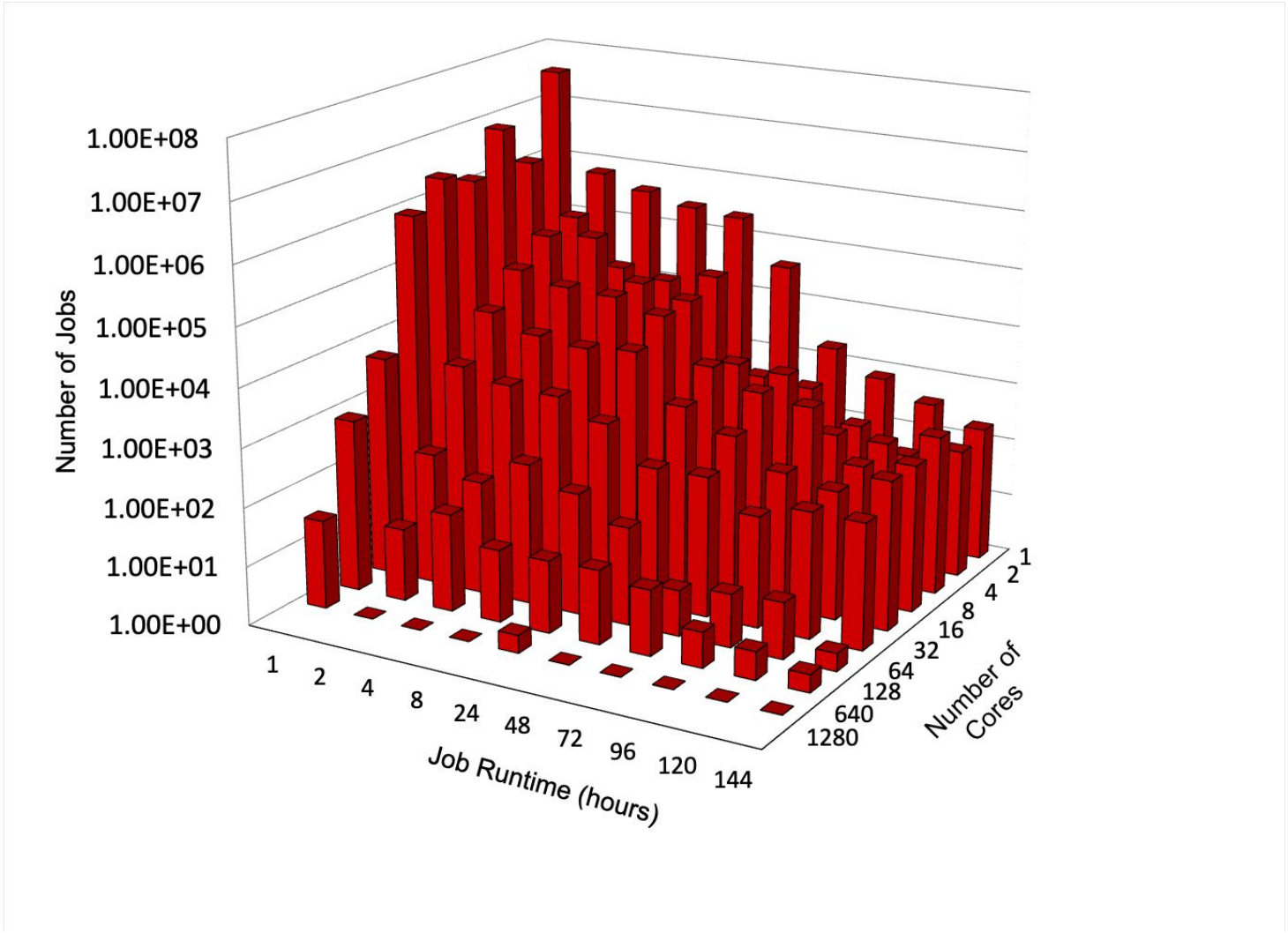
Accounts	April.1 2024- Sep. 30, 2024	Oct.1 2024- Mar. 31, 2025
Number of Active Users	1,002	977
Number of Total Users	4,447	4,675
Number of Project Groups	568 (428 active)	617 (445 active)
Storage		
High-Speed Storage Used (Arion)	19.5 PB (62% utilization)	21.6 PB (68% utilization)
Archival Storage Used	19.7 PB	20.5 PB
Compute		
Number of Jobs Run	29,225,632	52,719,740
Core-Hours Utilized	82,137,717 hrs	78,230,911 hrs*
System		
Number of Maintenance Sessions	No Preventative Maintenance (99.6% uptime)	98% Uptime

* Didn’t count all the retiring nodes during the transition for a month

Jobs and Compute Cores By Partition

Compute	# Jobs	CPU-hours	Utilization
Chimera	38,493,981	49,664,845	82.4%
Old Hi-Memory Nodes	4,553,928	6,267,942	100%
CATS	7,916,931	11,706,802	80.8%
GPU Nodes	1,580,180	9,342,661	45.8%
BODE2 (Retired in Stages)	174,720	1,248,661	NA
Total	52,719,740	77,699,801	76.0%

Job Mix


1.00E+08
1.00E+07
1.00E+06
1.00E+05
1.00E+04
1.00E+03
1.00E+02
1.00E+01
1.00E+00
Number of Jobs
1
2
4
8
24
48
72
96
120
144
Job Runtime (hours)
1
2
4
8
16
32
64
128
640
1280
Number of Cores

Top 10 Users Compute Core Hours

PI	Department	# Core-Hours	# Jobs
Nadkarni, Girish	Medicine	11,296,887	1,312,169
Raj, Towfique	Genetics and Genomic Sciences	9,002,533	7,437,704
Roussos, Panagiotis	Psychiatry	5,314,603	15,885,917
Zhang, Bin	Genetics and Genomic Sciences	4,793,963	1,659,271
Sharp, Andrew	Genetics and Genomic Sciences	4,047,520	1,004,870
Buxbaum, Joseph	Psychiatry	3,889,142	8,851,107
Pejaver, Vikas	Institute for Genomic Health	3,551,468	648,354
Kenny, Eimear	Institute for Genomic Health	3,500,271	4,204,697
Goate, Alison	Genetics and Genomic Sciences	2,878,032	440,827
Reva, Boris	Genetics and Genomic Sciences	2,501,294	1,966,540

Top 10 PIs GPFS High Speed Storage

PI	Department	Storage Usage(TiB)
Raj, Towfique	Genetics and Genomic Sciences	1,227
Roussos, Panagiotis	Psychiatry	1,140
Sebra, Robert	Genetics and Genomic Sciences	1,050
Nadkarni, Girish	Medicine	1,040
Charney, Alexander	Genetics and Genomic Sciences	998
Zhang, Bin	Genetics and Genomic Sciences	807
Sealfon, Stuart	Neurology	681
Goate, Alison	Genetics and Genomic Sciences	618
Pejaver, Vikas	Institute for Genomic Health	486
Kenny, Eimear	Institute for Genomic Health	454

Top Compute and Storage Usage Department/Institute

Department/Institute	Compute Core Hours
Genetics and Genomic Sciences	32,732,908
Psychiatry	11,282,033
Neuroscience	9,820,742
Institute for Genomic Health	7,134,205
Oncological Sciences	3,609,516
Medicine	3,557,949
Structural and Chemical Biology	2,861,128
AI and Human Health	1,585,380
Pharmacology	1,268,661
Pharmacological Sciences	1,211,748

Department/Institute	Storage (TiB)
Genetics and Genomic Sciences	6,297
Psychiatry	1,936
Neuroscience	1,631
Medicine	1,505
Oncological Sciences	1,318
Institute for Genomic Health	974
Neurology	764
Microbiology	325
AI and Human Health	204
Pathology	191

Top 10 PIs - GPU Usage Hours

PI	Department	GPU Hours	# Jobs
Raj, Towfique	Genetics and Genomic Sciences	1,812,364	104,141
Campanella, Gabriele	AI and Human Health	1,369,457	210,660
Shi, Yi	Pharmacological Sciences	1,204,023	1,180,859
Filizola, Marta	Structural and Chemical Biology	1,040,146	3,762
Nadkarni, Girish	Medicine	1,025,269	17,210
Schlessinger, Avner	Pharmacology	308,103	7,467
Tsankov, Alexander	Genetics and Genomic Sciences	241,408	3,817
Roussos, Panagiotis	Psychiatry	201,297	1,718
Shen, Li	Neuroscience	197,669	911
Davies, Terry	Medicine	192,129	185

Sumit: Total TSM Archival Storage Usage (Oct 2024- Mar 2025)

Current Archive Storage Usage	
Archived Data	20.5 PB
Number of Tapes Used	7,485

Statistics					
	Apr-Sept 2024	Oct 2024-Mar 2025		Apr-Sept 2024	Oct 2024-Mar 2025
Amount of Archived Data	1,400 TB	2,656 TB	Amount of Retrieved Data	335 TB	803TB
# of Users Who Have Issued Archive Commands	57	58	# of Users Who Have Issued Retrieve Operations	41	31

Minerva Publications > 1,900 Since 2012

We collect publications twice a year (Jan & May). Thank you!!!

Please cite Minerva platform and the technologies adopted in all your publications using:

- *Kovatch P, Gai L, Cho HM, Fluder E, Jiang D. Optimizing High-Performance Computing Systems for Biomedical Workloads. IEEE Int Symp Parallel Distrib Process Workshops Phd Forum. 2020 May;2020:183-192. doi: 10.1109/ipdpsw50202.2020.00040. Epub 2020 Jul 28. PMID: 33088611; PMCID: PMC7575271*

All publications utilizing Minerva resources must include one of the following acknowledgements depending on your funding status:

- **If you *have NIH-funded projects* on Minerva, you must include the following in all your publications:** "*This work was supported in part through the Minerva computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.*"

Year	# Pubs
2012	**55**
2013	**59**
2014	**62**
2015	**115**
2016	**149**
2017	**165**
2018	**133**
2019	**178**
2020	**146**
2021	**234**
2022	**174**
2023	**239**
2024	**214**
2025	**38**

To associate the CTSA grant UL1TR004419 or the CATS grant *S10OD030463* to an existing publication, please follow these instructions from the NIH (see the section "Associating Funding to your Publications").

Data Ark Data Usage Summary
Frequently Used Datasets Nov 2024 - Apr 2025

- Low user activities for recent 6 months.
- Other datasets show no user activity.

Dataset	Size (GB)	# of User Granted Access	# of Unique Users	# of Times Data Accessed	Users
Digital Pathology Slides (announced Dec. 2024)	1.5 PB	18	2	29,688,748	Gabriele Campanella Shengjia Chen
CBIPM-BioMe	4,608	44	3	363	Alanna Cote Kathleen Muenzen Robert Klein
GTEX	1,888	10	1	313	Ryan Conrad Thompson
1000 Genome	143	20	2	219	Gabriel Hoffman Laura Gail Sloofman
Reference Genome	142	12	2	13	Alanna Cote Ikjot Sidhu
MSDW OMOP	3,076	10	2	10	Alex Gallinat Daniel Jordan

# of unique active users: **11**
# of support tickets (including tickets automatically generated from user DUA submissions): **160**

Appendix II:
2024 User Survey Results

2024 Minerva Survey Results

We asked 5 questions:

Q1: Overall, how satisfied are you with LSF queue structure, storage, and compute resources (file system, GPUs, high-memory nodes, etc)?

Q2: Please rate current software environment (packages and services such as database, data transfer, container etc).

Q3: Please rate your satisfaction with operations (ticket system, responsiveness of staff, documentation, user support, etc).

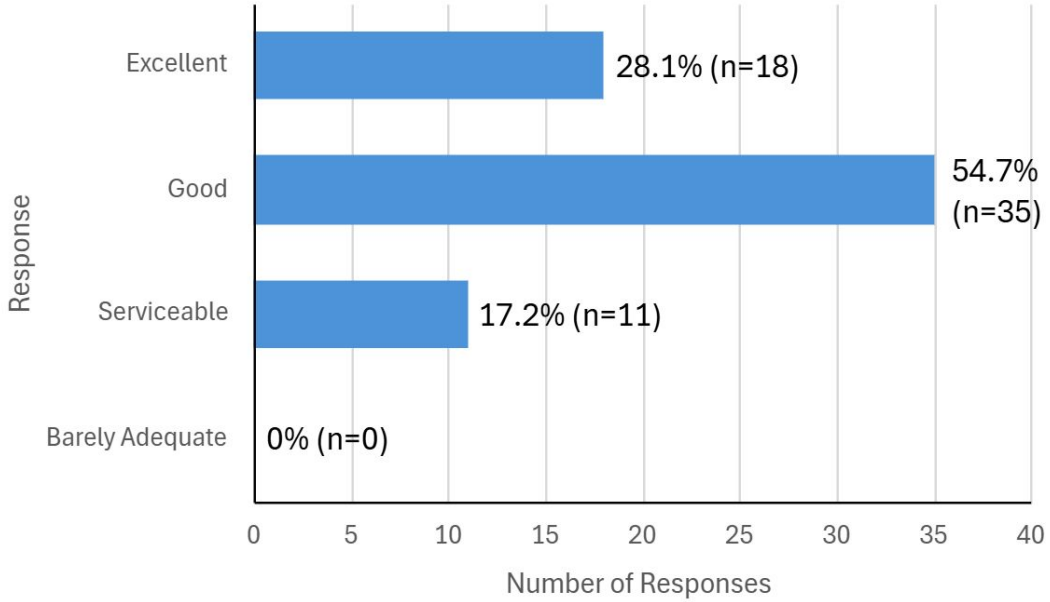
Q4: Is there any other new training session you would prefer adding to 2025 Minerva HPC training serials?

Q5: What suggestions do you have for improving our service?

We received 65 responses and 37 comments from 1,100 active users in Jan 2025 (6.0% response rate).

2024 Survey Results Question 1

Q1: Overall, how satisfied are you with LSF queue structure, storage, and compute resources (file system, GPUs, high-memory nodes, etc)?



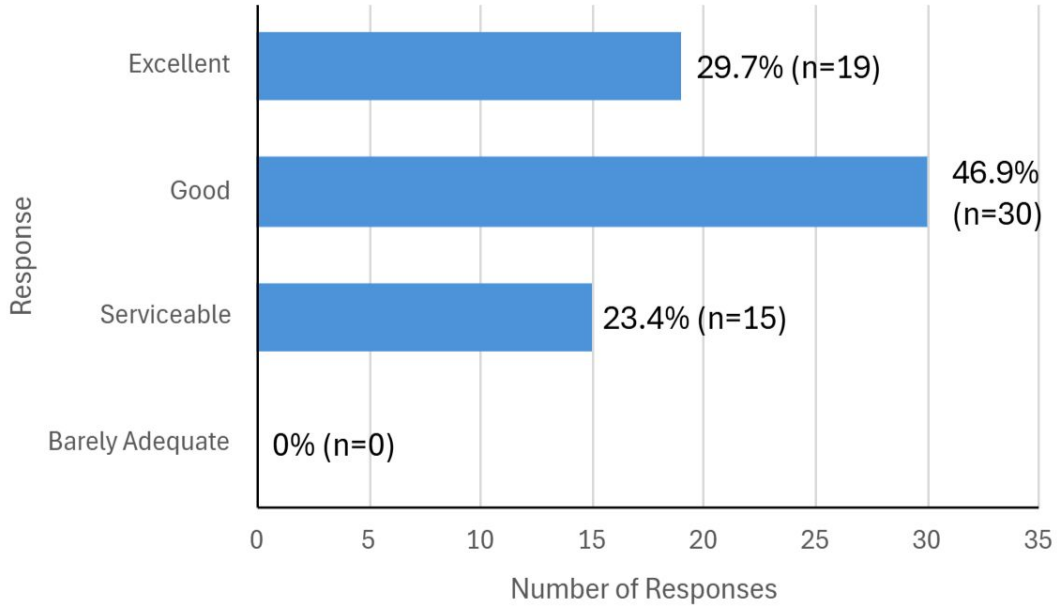
User satisfaction(>=Good)
2024: 83%
2023: 81%
2022: 84%
2021: 85%
2020: 81%
2019: 65%
2018: 54%

Raw comments (6): LSF queuing

- Queue times can be quite long. I prefer SLURM over LSF for job scheduling efficiency.
- The queue priority could be improved as usually a few users take most of the resources available.
- Recently improved but previously had times when I would wait hours in queue and never connect in the fall of 2024.

2024 Survey Results Question 2

Q2: Please rate current software environment (packages and services such as database, data transfer, container etc).



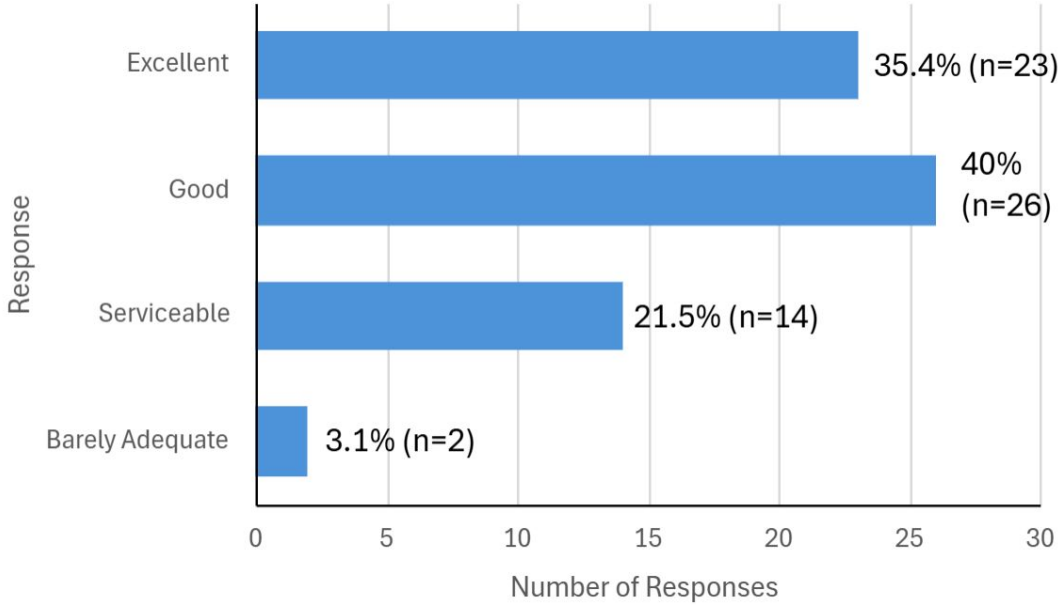
User satisfaction(>=Good)
2024: 77%
2023: 85%
2022: 93%
2021: 81%
2020: 80%
2019: 80%
2018: 67%

Raw comments(6): Software/Packages

- Some compatibility issues after the major hardware upgrade and change to Rocky9.
- Can be improved. The R package installation system is still a bit difficult for some packages and there are incompatibilities with the new LSF migration.
- Running python is honestly a huge hassle. Every time I need to install a new package, it feels like everything breaks. Maybe it would be possible to provide a written guide to this, if venv vs conda is recommended, etc.

2024 Survey Results Question 3

Q3: Please rate your satisfaction with operations (ticket system, responsiveness of staff, documentation, user support, etc).



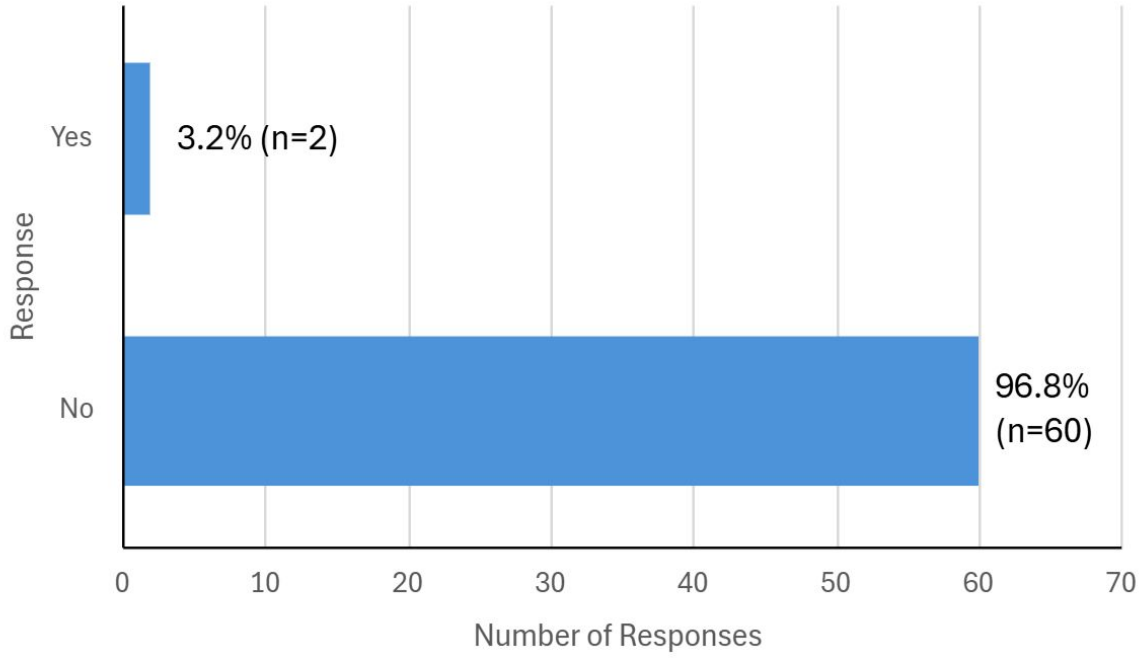
User satisfaction(>=Good)
2024: 75%
2023: 82%
2022: 88%
2021: 86%
2020: 91%
2019: 73%
2018: 80%

Raw comments (12):**Ticket System/Documentation**

- Sometimes ticket response time is way too long. Please hire more people. The staff is great, but it would be helpful to receive faster responses to the tickets.
- My main complaints are all due to a lack of staff/resources. I would prefer more documentation, and the user experience is pretty bare bones compared to HPCs I've used at other institutions.
- Also, a new ticket system where it is possible to follow up on a question would be helpful. Right now, each email opens a new ticket which can be confusing.

2024 Survey Results Question 4

Q4: Is there any other new training session you would prefer adding to 2025 Minerva HPC training serials?



Q4 *(Continued)*: If **yes**, please specify any new training session topic you are interested in. Currently, the following sessions are available:

1. Minerva Introduction.

2. LSF Job Scheduler.

3. Open OnDemand.

4. Introduction to GPU resources.

5. Accelerating Biomedical Data Science with GPUs: Practical Approaches and Tools.

6. Leveraging Large Language Models in Biomedical Research.

7. How to Accelerate Genome Analysis Toolkit (GATK) by using Parabricks.

8. Data Ark.

- No responses.

2024 Survey Results: More Raw Comments

Thank you for your feedback! We are posting all the responses on our website.

Positive Comments:

- Love working with the HPC team and for the most part, the services offered are great. Minor hiccups related to the substantial update to the system.
- Everything about Minerva and the team is amazing!
- As someone who works/has worked with multiple different HPCs, I find the Minerva management to be excellent. Documentation is good and support is helpful, making it simple for a moderately experienced researcher to make use of the service.
- Year after year, HPC has always been able to help me with whatever needs I had and offered computational resources in a timely manner. Thank you!

Other comments:

- Maybe more beginner friendly. Also work with MSSM VPN to develop a more feasible and easier way to use the login method. The VPN doesn't work outside of the US, and there is no way to connect to HPC when you are traveling.
- I think monitoring user activity can be improved. The only issue that is affecting the quality of HPC and our experience is how other users seem to be utilizing these resources. And I think that issue can be fixed by communication and activity monitoring. Sometimes I cannot run a simple submit command but could see in 'top' results that other users are running memory and CPU heavy tasks on the login node.
- Using HIMEM nodes when the tasks do not require large amounts of memory is an issue; or using GPU nodes when the tasks do not use GPUs at all simply due to these nodes having "better" hardware.

Appendix III:
2024 Data Ark User Survey Results

2024 Data Ark Survey Results

We asked 4 questions:

Q1: How would you rate your satisfaction with the Data Ark: Data Commons quality and availability at Mount Sinai?

Q2: How would you rate your satisfaction with Data Ark support?

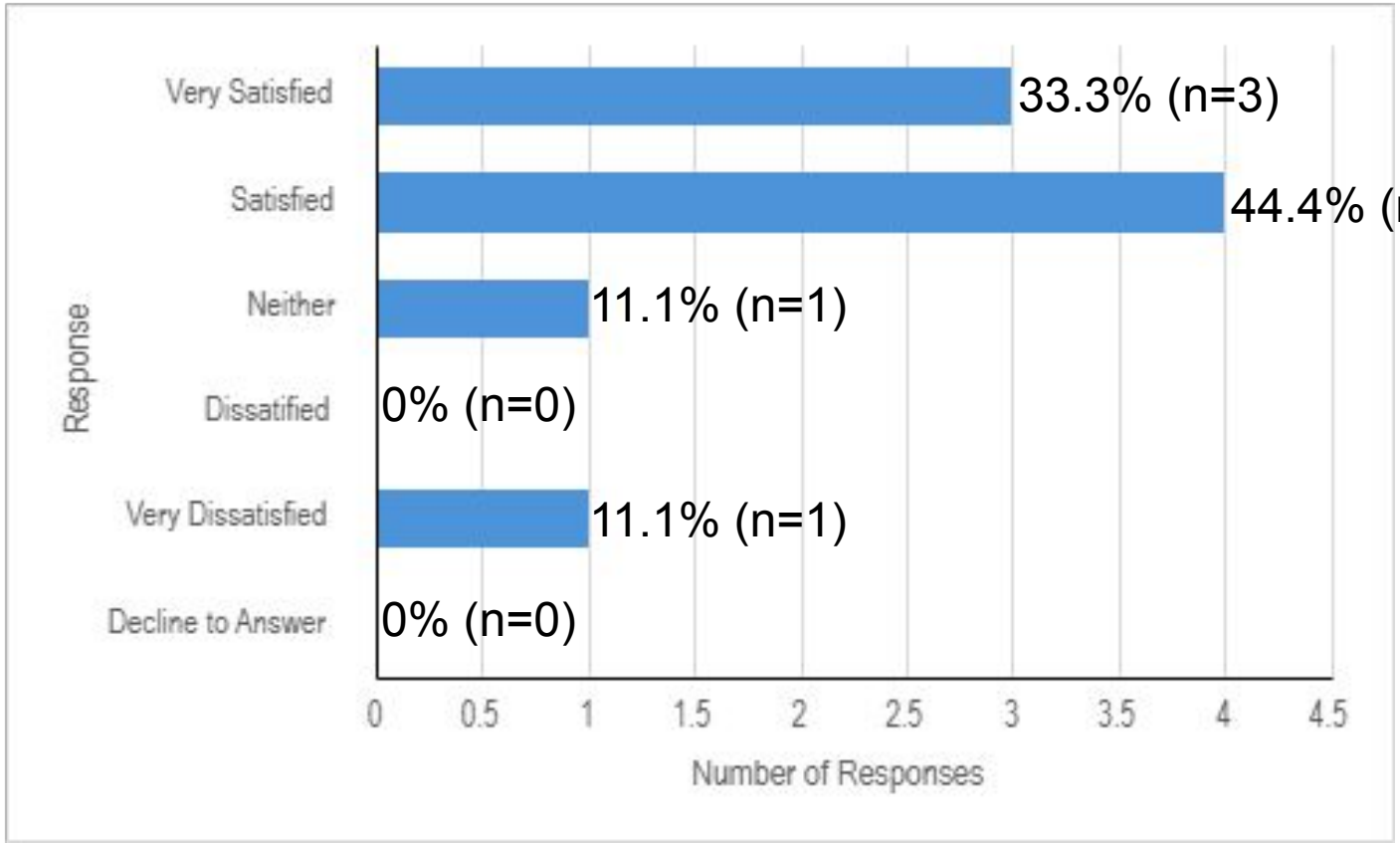
Q3: What barriers exist preventing your usage of Data Ark?

Q4: Please enter any other feedback on Mount Sinai's Data Ark: Data Commons including suggestions for data sets, availability, accessibility, support, new features and development.

We received 9 responses and 2 comments from 56 active users in Jan 2025 (16.0% response rate).

2024 Survey Results Question 1

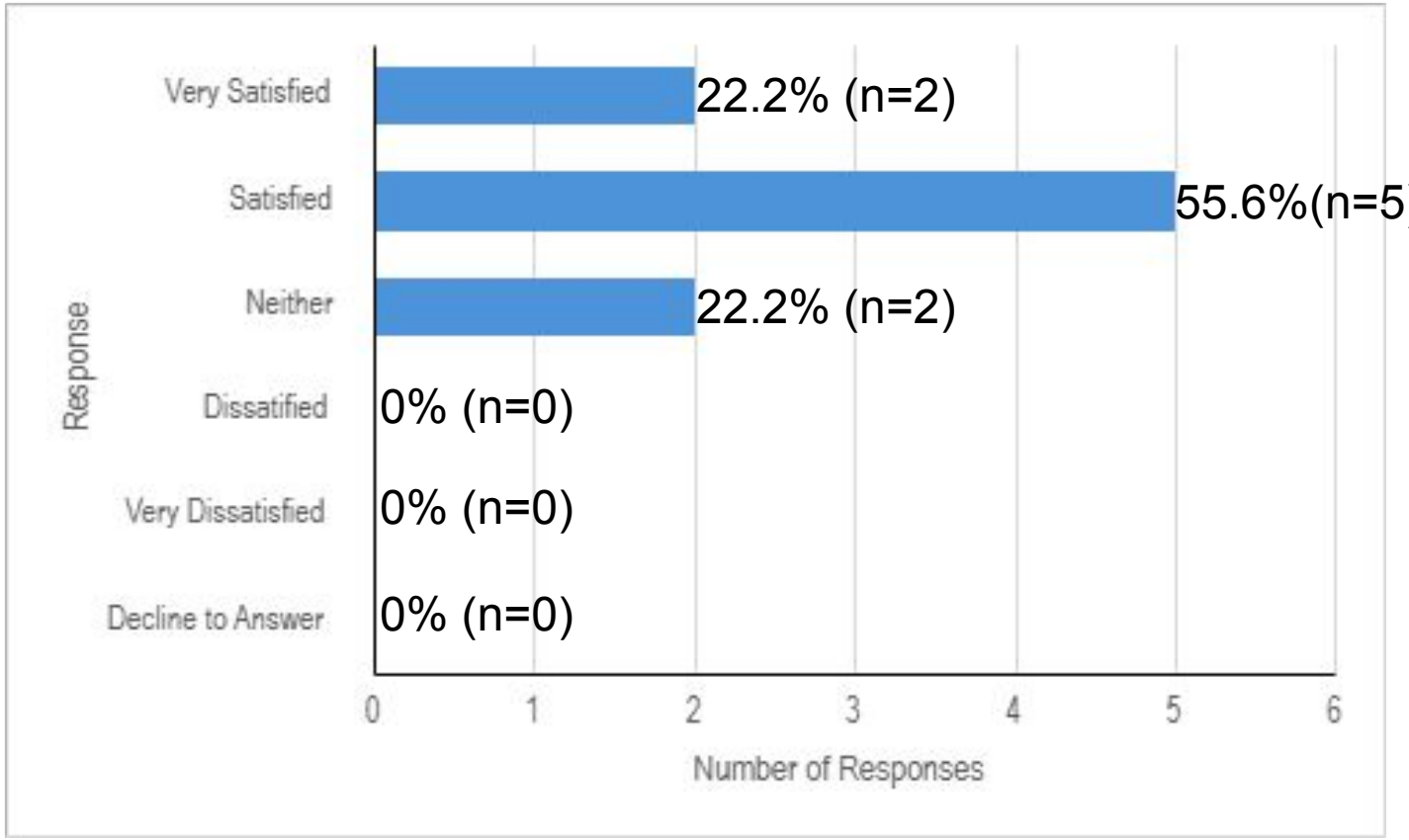
Q1: How would you rate your satisfaction with the Data Ark: Data Commons quality and availability at Mount Sinai?



User satisfaction(>=Satisfied)
2024: 78%
2023: 80%
2022: 72%

2024 Survey Results Question 2

Q2: How would you rate your satisfaction with Data Ark support?



User satisfaction(>=Satisfied)
2024: 78%
2023: 84%
2022: 72%

2024 Survey Results Question 3

Q3: What barriers exist preventing your usage of Data Ark?

Comment	Response
• Having to request approval individually for each dataset is annoying. I understand why it's necessary for datasets with access requirements, but it seems silly for datasets with public unrestricted access.	• Thank you for the comment. Users are required to review and adhere to dataset-specific requirements, including the proper citation of corresponding publications. • While individual Data Use Agreements (DUAs) are necessary, we will streamline the DUA completion process for public unrestricted datasets. By caching user-entered information, we may eliminate the need for repeated user entry, making the process more user-friendly.

2024 Survey Results Question 4

Q4: Please enter any other feedback on Mount Sinai's Data Ark: Data Commons including suggestions for data sets, availability, accessibility, support, new features and development.

Comment	Response
•There are 4 public datasets that I remember wishing were in Data Ark: 1. hg18 reference. 2. dbSNP vcf(s) (https://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/). 3. dbNSFP (https://www.dbnsfp.org/download). 4. multiple sequence alignments from Zoonomia project (https://zoonomiaproject.org/the-data/).	•Thank you for your feedback. We will look into the possibilities of adding the requested data sets to Data Ark and improving accessibility.