Shared Research Computing Policy Advisory Committee (SRCPAC)

Spring 2023 Meeting

Chris Marianetti, *Chair of SRCPAC*
Alexander Urban, *Co-Chair of SRCPAC*

SRCPAC Agenda

Welcome & Introductions
- Chris Marianetti, Chair of SRCPAC
- Alexander Urban, Co-Chair of SRCPAC

Long Term Strategic Thinking about University Needs for Computing and Storage
- Jeannette Wing, EVP for Research

High-Performance Computing Update
- Kyle Mandli, Chair of the HPC Operating Committee
- John Villa, Manager of High Performance Computing, CUIT

Research Computing Services Update
- Axinia Radeva, Manager of CUIT Research Services

Data Catalog Project and the POC with school of Nursing/Newsletter for Researchers
- Maneesha Aggarwal, CUIT AVP, Academic, Emerging Technologies & Research Services

Foundations for Research Computing Update
- Marc Spiegelman, Chair of the Foundations for Research Computing Advisory Committee
- Jeremiah Trinidad-Christensen, Head of Research Data Services, Columbia University Libraries

Other Business & Closing Remarks
- Chris Marianetti, Chair of SRCPAC
- Alexander Urban, Co-Chair of SRCPAC

Long-Term Strategic Thinking about University Needs for Computing and Storage

Jeannette Wing, *Executive Vice President for Research*

Charge

The Office of the EVP for Research and CUIT charges this committee: **to recommend a strategic plan for the University's future computational and data infrastructure for research**. The committee should consider all the major elements of this plan, including but not limited to:

- Compute resources
- Data resources for analysis, sharing, storage, archiving, privacy and security
- Technology skills required
- Policy impacts
- High level cost implications

The committee should provide its recommendations by the end of Academic Year 2023-24 (May 31, 2024).

Faculty Committee

The faculty committee consists of a diversity of users and needs from schools and institutes across the University, with membership as follows:

- **Timothy Berkelbach**, Associate Professor of Chemistry; Faculty of Arts and Sciences
- **Roisin Commane**, Assistant Professor of Earth and Environmental Sciences, Atmospheric Composition Group, Lamont Doherty Earth Observatory; Faculty of Arts and Sciences
- **Wojciech Kopczuk**, Professor of Economics and of International and Public Affairs; Faculty of Arts and Sciences and School of International and Public Affairs
- ***Hod Lipson, co-chair***, James and Sally Scapa Professor of Innovation in the Department of Mechanical Engineering; Co-Director, Maker-Space Facility; School of Engineering and Applied Sciences
- **Ciamac Moallemi**, William von Mueffling Professor of Business; Columbia Business School
- ***Darcy Peterka, co-chair***, Senior Research Scientist; Scientific Director of Cellular Imaging; Mortimer B. Zuckerman Mind Brain Behavior Institute
- **Muredach Reilly** (interim), Florence and Herbert Irving Endowed Professor of Medicine; Director, Irving Institute for Clinical and Translational Research; Associate Dean for Clinical and Translational Research; Columbia University Irving Medical Center

Faculty Committee (cont.)

Ex-officio members are:

- **Robert Cartolano**, Associate Vice President for Technology and Preservation, Columbia University Libraries
- **Gaspare LoDuca**, Chief Information Officer and Vice President for Information Technology, Columbia University Information and Technology
- **Alexander Urban**, Assistant Professor of Chemical Engineering, School of Engineering and Applied Sciences and Chair of the Columbia Shared Research Computing Policy Advisory Committee (SRCPAC)
- **Jeannette M Wing**, Executive Vice President for Research; Professor of Computer Science, Office of Research

Staffing the committee will be:

- **Maneesha Aggarwal**, Assistant Vice President, Academic, Emerging Technologies & Research Services, Columbia University Information and Technology
- **Sophie Thuault-Restituito**, Chief of Staff and Executive Director for Special Projects, Office of Research

High Performance Computing Updates

Kyle Mandli, *Chair, Research Computing Operating Committee*
John Villa, *Manager, High Performance Computing, CUIT*

HPC Governance

- HPC operations are governed by the faculty-led **HPC Operating Committee**, chaired by **Kyle Mandli**.
- The operating committee reports to SRCPAC and reviews business and usage rules in open, semiannual meetings
- The last meeting was held on March 26, 2020 and the next one will be in Winter 2022 to discuss consolidation strategy.
- **All HPC Users (Ginsburg, Terremoto, Habanero) are invited to participate.**

Habanero Retirement

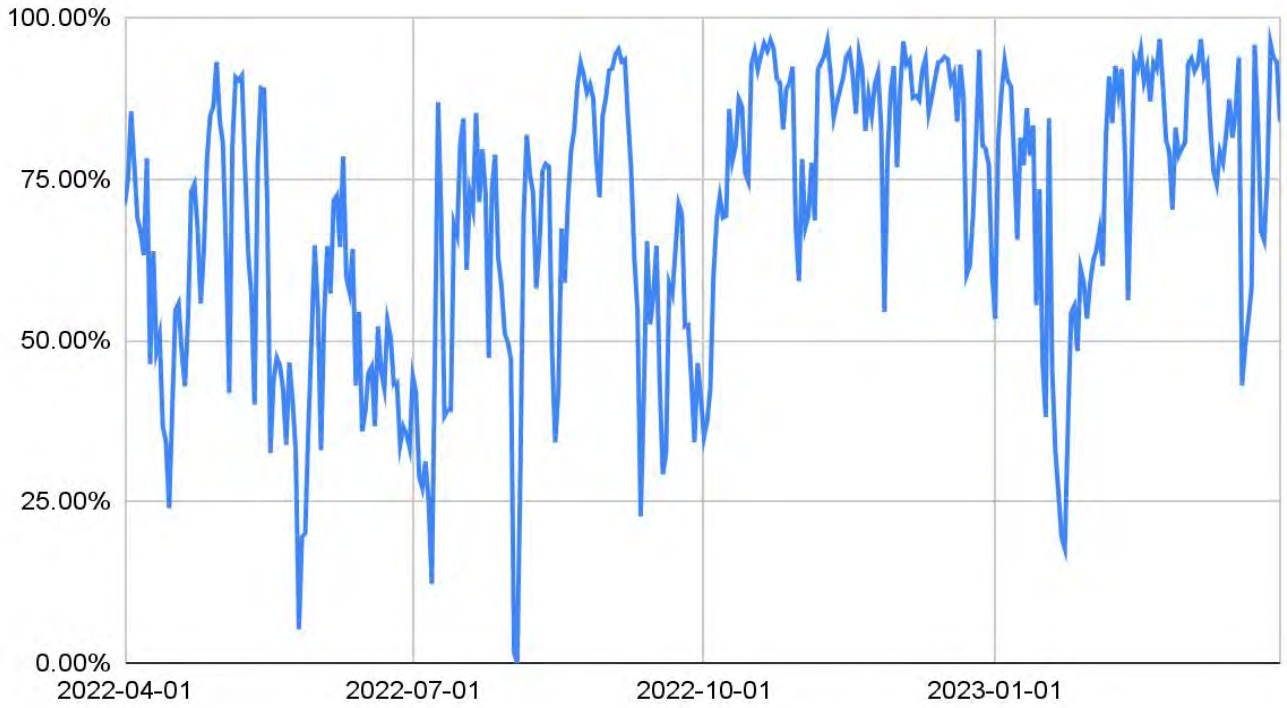
- The remainder of Habanero has been moved out of the high density racks within the Data Center
- Phase 2 will retire in May 1, 2023.
- The Phase 2 users will be move to the free tier, which is a low priority queue.
- Although storage is still accessible, all users have been advised to move their data off, as the storage is end-of-life and no longer supported.

TERREMOTO

- Launched in December 2018
- Expanded in December 2019
- 5 year lifetime

- **Phase 1 retirement - December 2023**
- **Phase 2 retirement - December 2024**

Terremoto - Cluster Usage in Core Hours



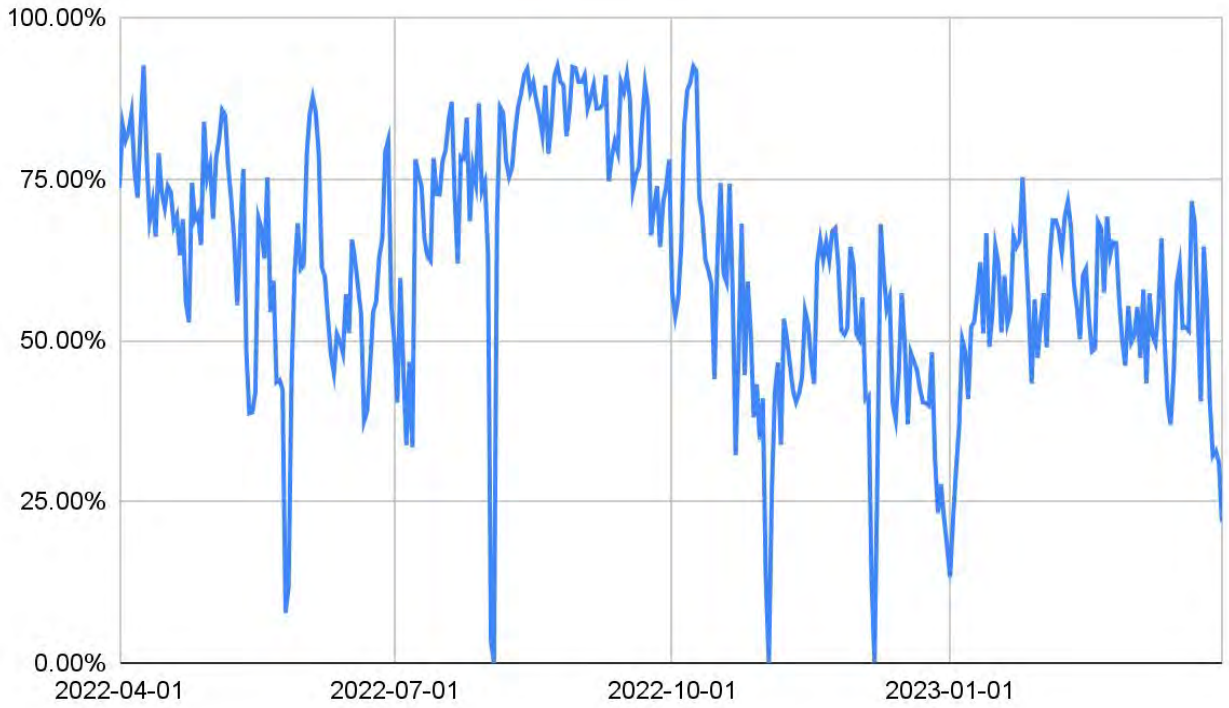
Total core hours in the past 12 months = ***20 million***

Ginsburg

- Ginsburg Expansion 2 (Phase 3) went live in December 2022
- Ginsburg now cluster total to **286 nodes, 9152 cores and 39 GPU hardware accelerated systems.**

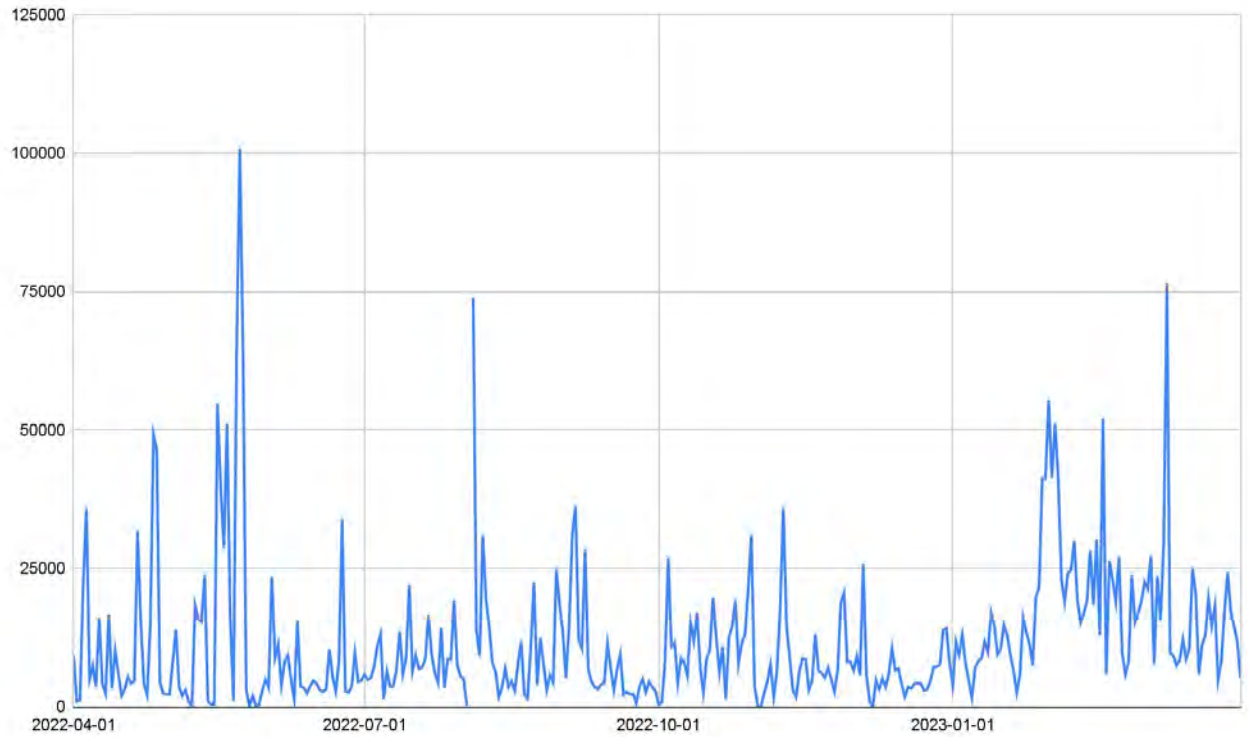
Ginsburg Phase 1 retirement - 2025
Ginsburg Phase 2 retirement - 2026
Ginsburg Phase 3 retirement - 2027

Ginsburg - Cluster Usage in Core Hours



Total core hours the past 12 months = ***43 million***

Ginsburg - Daily Job Count



GPU Cluster - Manitou

- Manitou was delivered in late February 2023 and is currently live.
- Manitou Phase 2 expansion is experiencing supply chain issues.

- The cluster will total 15 nodes when completed.
  - 13 nodes with 1TB of memory 96 cores and 8 A6000 GPUs with NVLink
  - 2 nodes with 256G of memory 32 cores and 4 A6000 GPUs

RFQ Committee

RFQ Committee convened early March
Attendees: Kyle Mandli, Julia Hirschberg, Bob Mawhinney, Rob Lane

- Specifications were agreed upon
- RFQ was sent to 15 vendors
- Responses just received and reviewed
- The finalists for the final round have been narrowed down to
  - Penguin Systems
  - HPE
  - Dell
  - Lenovo
  - Aspen/Panasas
- Open order period should start May 15 - June 15
- Menu and final pricing to be distributed prior

RFQ Committee

- Our Vision for the next phase of HPC
  - Centralized Storage
  - Centralized Provisioning
  - Centralized Service

HPC Support Services

- **Email**
  - rcs@columbia.edu - general questions
  - hpc-support@columbia.edu - HPC technical questions
- **Office Hours (Online)**
  - Speak with HPC support staff via Zoom from 3pm – 5pm on 1st Monday of month: Registration required
- **Group Information Sessions**
  - HPC support staff meet with your group, upon request
- **Training Workshops every semester (Online)**
  - Introduction to Linux
  - Introduction to Scripting
  - Introduction to High Performance Computing
- **Cloud Computing Consulting**
  - Complimentary assistance moving HPC workloads to the cloud

Research Computing Services Updates

Axinia Radeva, *Manager, Research Computing Services, CUIT*

Research Computing Services

- **Research Computing Services (RCS) Goals**
  - Expand RCS portfolio
  - Support research activities across Columbia University
    - Provide faster processing times
    - Increase productivity
    - Improve accuracy
    - Foster greater collaboration
    - Achieve cost savings for researchers

Current Research Computing Services

Embedded Research Computing Support
We provide embedded research computing support to CPRC, SSW, DSI, Psych, and other affiliates on the Morningside and Medical Center campuses.

Secure Data Enclave (SDE)
A virtual platform used for working with **secure data sets**.

Electronic Research Notebooks with LabArchives
An online platform specialized in **organizing and storing laboratory data**, as well as enabling information sharing and collaboration, all with automated backups and a comprehensive audit trail. Enterprise license is covered by CUIT and the Libraries.

Globus
Our enterprise Globus subscription helps you efficiently, securely, and reliably transfer data directly between systems, including between HPC clusters and Amazon S3, Google Drive, Box and more.

Cloud Research Computing Consulting
Looking to utilize the Cloud to further your research efforts? Our team can help you determine the best resources and configurations to support your needs and assist with onboarding.

Access National HPC Campus Contact
Columbia researchers can try out the **Columbia's Discover allocation** and receive guidance for applying for free Access national HPC resources.

SnapGene license discount
A molecular biology software that allows users to plan, visualize, and document molecular biology procedures. CUIT offers the opportunity to purchase an annual SnapGene license at a **reduced price** through the University's bulk license.

Research Computing Services

- **Research Computing Services Update**
  - Overleaf
  - Embedded team
  - Secure Data Enclave hardware upgrade
  - Globus - Connector Open Access
  - XSEDE/ACCESS
  - LabArchives
  - SnapGene

Research Computing Services Update

Overleaf Professional – Coming July 2023!

- Online LaTeX and Rich Text collaborative writing and publishing tool that facilitates the writing, editing and publishing of scientific documents
- **CUIT and the Libraries are partnering** to provide an enterprise license for all Columbia users (students, faculty, and researchers)
- Columbia's ~11K existing Overleaf users will be able to transfer to the University license seamlessly
- Columbia's users often come from Computer Science, Physics, Economics, Electrical Engineering, and Mathematics

Research Computing Services Update

Embedded Research Computing Support

- Purpose:
  - Provide on-site research computing support to Columbia research departments or centers
  - Hired and trained by CUIT
  - Provide curated services depending on local needs
- 2023
  - Two new roles added to the RCS team to provide more resources to the embedded engineers and expand capabilities
    - Embedded Research Computing Specialist
    - Sr. Research Systems Engineer, Cloud and Security

Research Computing Services Update

Secure Data Enclave (SDE)

- Since 2018, SDE provides researchers with a virtual cold room to analyze and collaborate on projects with restricted data sets
- Hardware Upgrade 2023
  - Current hardware going out of support February 2024
  - More storage and computing resources needed to address growing demand
  - New blades will be usable until 2030, allowing costs to be spread over 6 years
  - New blades can support GPU cards if there is demand and resources

Globus

Highly recommended for high speed file transfers to/from the HPC clusters! CUIT maintains a University-wide subscription.

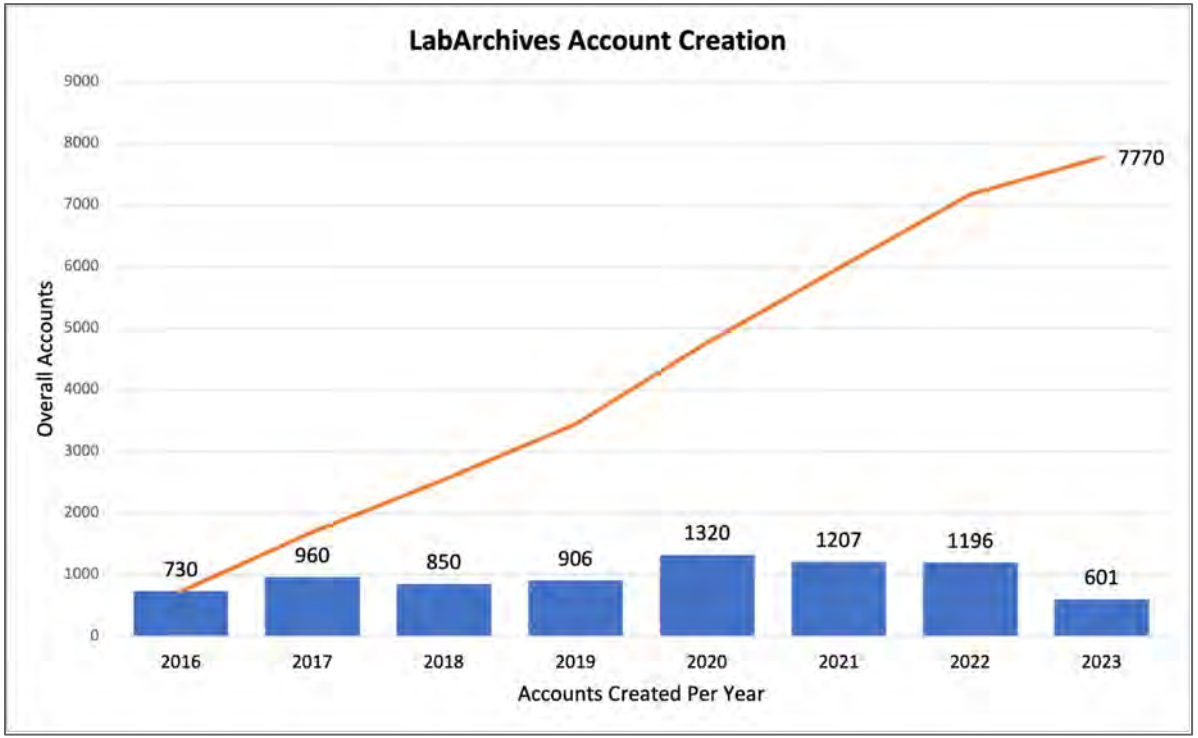
- **FLEXIBLE:** Transfer datasets of any size to/from Amazon S3, Google Drive, Box, and more!
- **FAST:** Quicker than SCP, and won't affect other users by clogging the login nodes
- **FREE:** Globus is provided at **no cost to you**, and it's easy to get an account – simply email globus@columbia.edu with your UNI, and we'll send you an account invitation
- **RELIABLE:** Transfers automatically resume after temporary network disconnections
- **COLLABORATIVE:** Globus allows users to share data with colleagues at other institutions

Research Computing Services Update

XSEDE/ACCESS

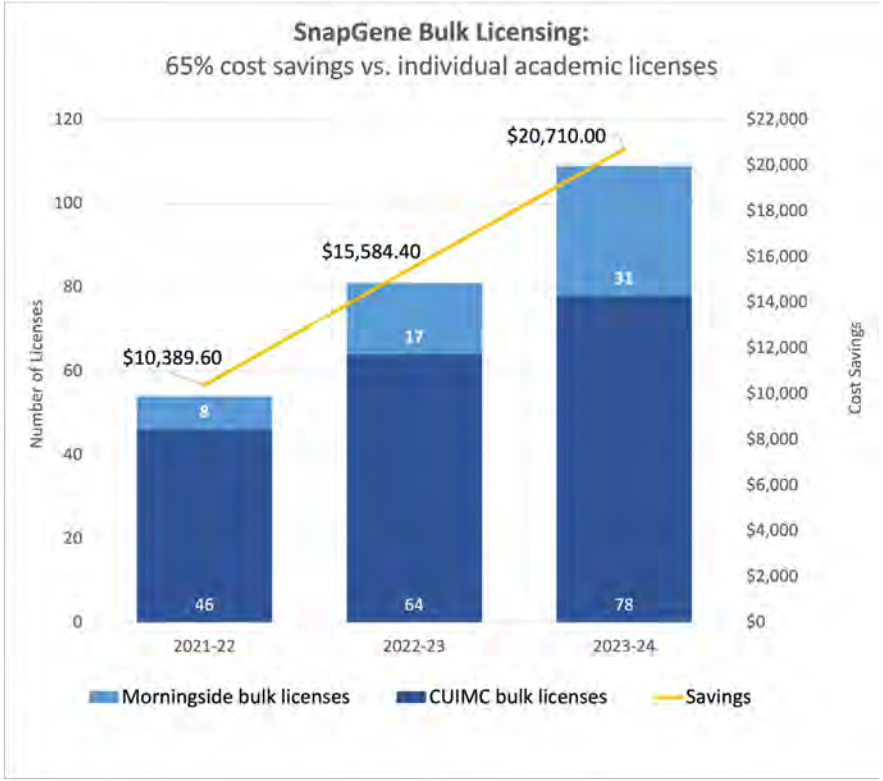
- **XSEDE (Extreme Science and Engineering Discovery Environment)** now known as **ACCESS (Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support)**, is an NSF-funded, nationwide collection of supercomputing systems available to researchers through merit-based allocations.
- CUIT RCS has five Campus Champions to assist you on obtaining an allocation
- Columbia's Discover allocation (used for test jobs) was renewed in February

2023 Electronic Research Notebooks



LabArchives Forums at Morningside and CUIMC were held in Fall 2022

SnapGene bulk licensing



- Doubled the number of users of SnapGene since Spring 2022
- A SnapGene webinar was held February 2023, offering Columbia users insights directly from a field scientist at SnapGene
  - Will hold webinars on semesterly basis

Top Users

SnapGene®
Software for molecular biology

- Chemistry
- Systems Biology
- Biological Sciences
- Genetics and Development
- Pathology and Cell Biology
- Biochemistry & Molecular Biophysics

Research Computing Services

Research Computing Services support is available to discuss your research technology needs by emailing rcs@columbia.edu.

Data Catalog Project and the POC with School of Nursing/Newsletter for Researchers

Maneesha Aggarwal, *CUIT AVP, Academic, Emerging Technologies & Research Services*

SRCPAC Spring 2023 Update

April 24, 2023

Foundations Mission

Foundations for Research Computing provides **informal training** for Columbia University graduate students and postdoctoral scholars to develop fundamental skills for harnessing computation: core languages and libraries, software development tools, best practices, and computational problem-solving.

Purpose: to provide the investment in people and computational skills required to compliment our investment in hardware, software and systems adminstration

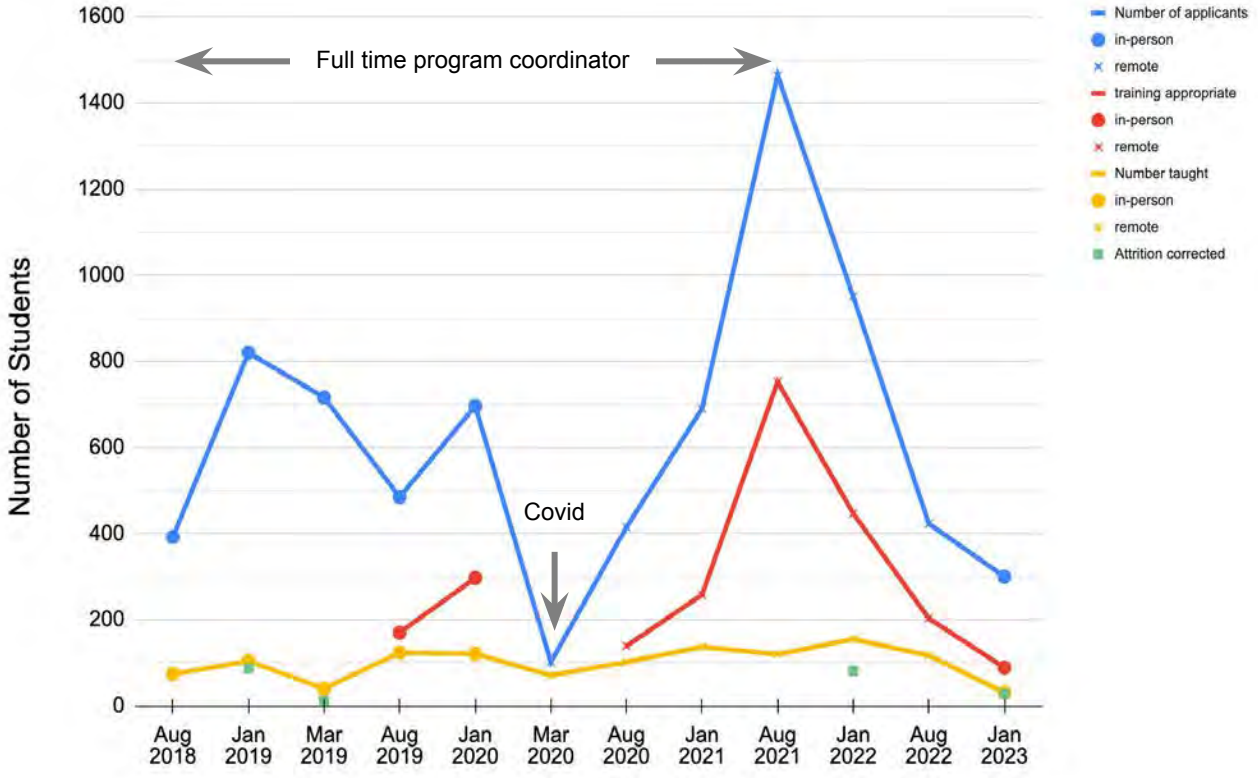
Foundations Primary Activities

- **Novice trainings**: 2 day training based on Software Carpentry curriculum for novice learners, learning Git, UNIX, and either R or Python
- **Data Club**: revamping of Python Users Group: twice-monthly meeting for those using computation in their research or interest about specific, more advanced topics
- **Intermediate intensives**: 1 day training for intermediate learners
- **Workshops**: 1.5 - 2 hour training opportunity to advance computational skills in a group setting. Workshops are often led by partners including CUIT and the Libraries

Novice Training Bootcamps

- 12 Bootcamps since Aug 2018 (2-3/year)
- Half were remote due to Covid – remote format presented challenges, particularly at Novice level
- Return to in-person, January 2023

Novice Training Data



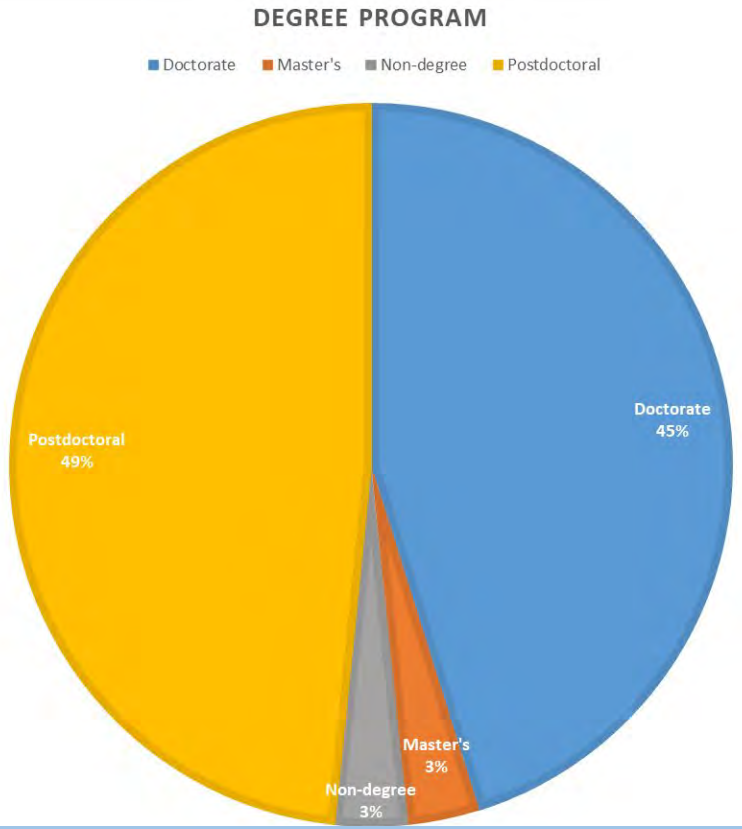
Some Observations

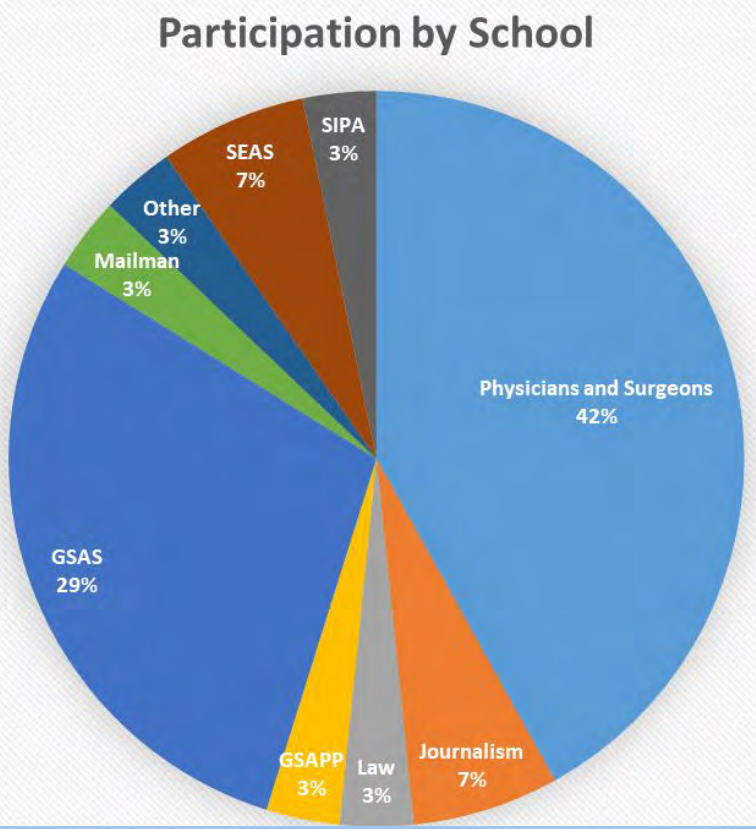
- Demand always exceeds supply
- Even when filtered for background.
- Novice training is extremely labor intensive – challenging to scale
- Identifies considerable demand for more advanced training
- All of this requires a full-time program coordinator

Spring 2023 In-person Novice Training

- Dates: January 12th & 13th
- Location: Uris Hall
- 300 applicants of all proficiency levels
- Selected 31 novice researchers
- Primarily Doctoral and Postdocs
- High retention factor over 2 days

Spring 2023 Novice Training (31 participants)



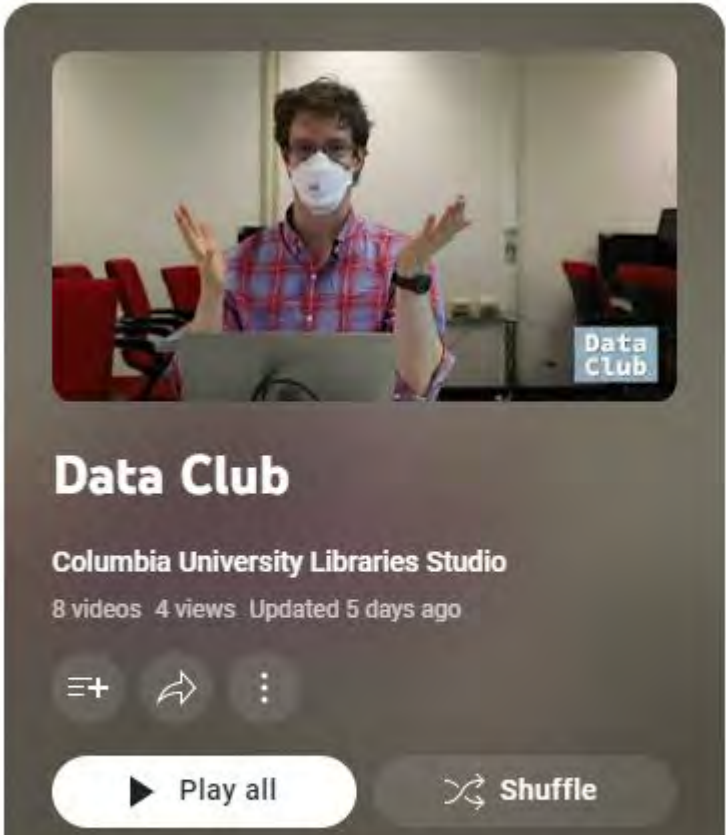


Data Club Sessions

Spring Sessions

Held every two weeks in Lehman Library
Recorded and made available on YouTube

- Intro to Python
- Intro to Pandas
- Extending Pandas with Dask
- Python and Relational Databases
- Exploratory Analysis of Textual Data
- Intermediate Textual Data Analysis
- XArray's Children



Foundations: Some Questions to be addressed

- **Who is the target audience?**
  - Started in SRCPAC to enhance “research computing” (mainly grad students/postdocs)
  - But demand is larger and Libraries supports the entire University
- **How to scale to meet demand?**
  - Is a model based on volunteer instructors and helpers sustainable?
  - How do we find, foster and cultivate high-quality instructors?
  - How do we address Novice training between bootcamps?
- **How do we choose content/level?**
  - Currently, Foundations focuses on Unix, Git, and either R or Python primarily at a novice level
  - How to stay nimble in the face of changing technology?
  - Can we identify appropriate intermediate level offerings with broad interest?
  - Can we develop better mechanisms for investigating/vetting these topics
- **What funding/structures would be needed to support any new ideas suggested for the above?**

Foundations: Issues for Moving forward

Foundations requires full-time personnel and oversight

- Planning for two new hires in the Libraries
  - **Computational Research Instruction** (CRI) **Librarian**
    - Develop and oversee the administrative and instructional matters related to the Foundations. e.g. recruitment and oversight of students, volunteers.
    - Engage with faculty and administrators for input into program assessment and development.
    - Coordinate with partners in CUIT, EVPR, and the Libraries to promote workshops offered by campus partners on topics related to data science, high-performance computing, and computational research
    - Cultivate and manage the community of trained instructors to plan and facilitate workshops, boot camps, and additional training opportunities at the novice and intermediate levels.
    - Represent CU at Carpentries
    - Oversee the management of community-facing and Foundations related support programs such as Data Club.
    - *Interviews scheduled for last week of April, with goal of starting in July*
  - **Director for Digital Scholarship**
    - Oversees the CRI librarian (and related positions)
    - *Hope to hire soon*
- **Also requires reinvigorated engagement with Faculty & other stakeholders**

Summary/Conclusions

- The need and rationale for Foundations has not changed
- But the mechanics/structure requires review with all stakeholders
- Now is particularly timely, given new potential hires
- SRCPAC should be a natural place to seek new leadership
- Happy to take any questions