Shared Research Computing Policy Advisory Committee (SRCPAC)

Fall 2022 Meeting

Chris Marianetti, *Chair of SRCPAC*

SRCPAC Agenda

Welcome & Introductions
- Chris Marianetti, Chair of SRCPAC

Data Center Capacity
- Halayn Hescock, Sr Director of CUIT Research Services

Long Term Strategic Thinking about University Needs for Computing and Storage
- Jeannette Wing, EVP for Research
- Gaspare LoDuca, CIO/VP for Information Technology, CUIT

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

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

Foundations for Research Computing Update
- Marc Spiegelman, Chair of the Foundations for Research Computing Advisory Committee
- Jonathan Cain, Associate University Librarian for Research and Learning

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

Data Center Capacity

Halayn Hescock, *Senior Director, Research Services, CUIT*

Data Center Capacity

Recap in brief

- 2010 NIH G20 Construction Grant
  - upgraded data center power to 900kW
  - 500kW is dedicated to HPC
  - 400kW (200kW redundant) for the remainder of the data center
- 2018 Upgrades
  - upgraded piping from chiller plant
  - installed 16 high density racks

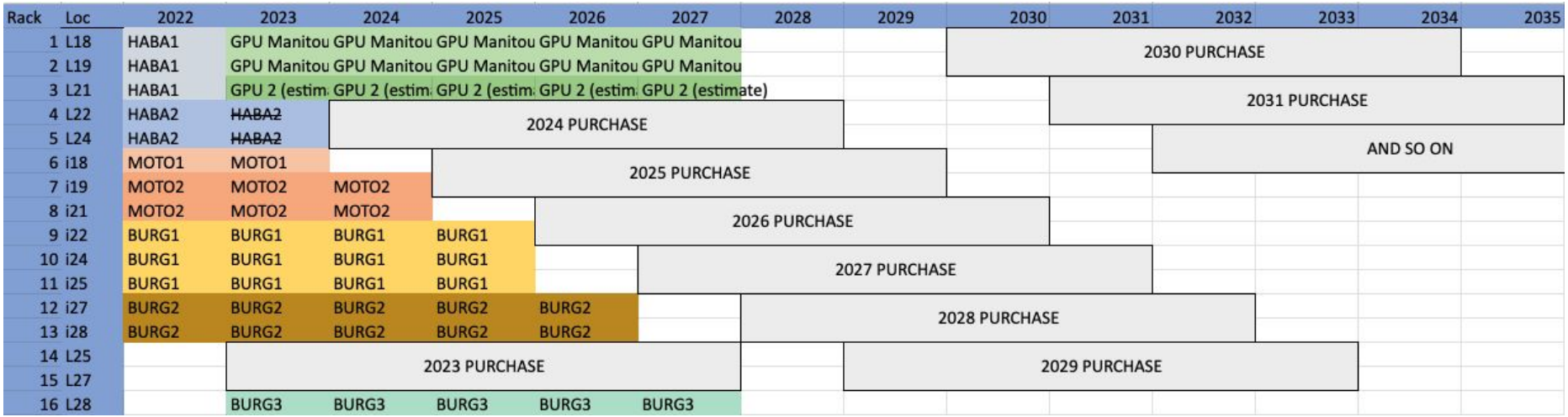
Data Center Capacity

- Data Center Capacity is based on three major factors
  - Cooling
  - Power
  - Space
- HD racks can run at ~25kW and if the 16 HD racks are fully loaded, we will be at 400kW
- We are currently occupying 13 of the 16 HD racks, using approximately 200kW
- But we can double our electrical capacity within the existing 16 HD racks
- Newer, denser equipment and strategic management of the hardware raise rack kW
- Retirement of hardware will keep racks rotating

Data Center Capacity

Rack	Loc	2022	2023	2024	2025	2026	2027	2028	AND SO ON...
1	L18	HABA1	GPU Manitou	GPU Manitou	GPU Manitou	GPU Manitou	GPU Manitou		
2	L19	HABA1	GPU Manitou	GPU Manitou	GPU Manitou	GPU Manitou	GPU Manitou		
3	L21	HABA1	GPU 2 (estimate)	GPU 2 (estimate)	GPU 2 (estimate)	GPU 2 (estimate)	GPU 2 (estimate)		
4	L22	HABA2	~~HABA2~~						
5	L24	HABA2	~~HABA2~~						
6	i18	MOTO1	MOTO1						
7	i19	MOTO2	MOTO2	MOTO2					
8	i21	MOTO2	MOTO2	MOTO2					
9	i22	BURG1	BURG1	BURG1	BURG1				
10	i24	BURG1	BURG1	BURG1	BURG1				
11	i25	BURG1	BURG1	BURG1	BURG1				
12	i27	BURG2	BURG2	BURG2	BURG2	BURG2			
13	i28	BURG2	BURG2	BURG2	BURG2	BURG2			
14	L25								
15	L27								
16	L28		BURG3	BURG3	BURG3	BURG3	BURG3		
	U7	FREE TIER nodes will be maintained outside of the HD Racks							
	U8	FREE TiER							
	U9	FREE TiER							
	U10	FREE TiER							

Data Center Capacity



Assuming steady state, filling two racks per year

Research Computing, Into the Future

Jeanette Wing, *Executive Vice President for Research*
Gaspare LoDuca, *Chief Information Officer and Vice President for Information Technology*

High Performance Computing Updates

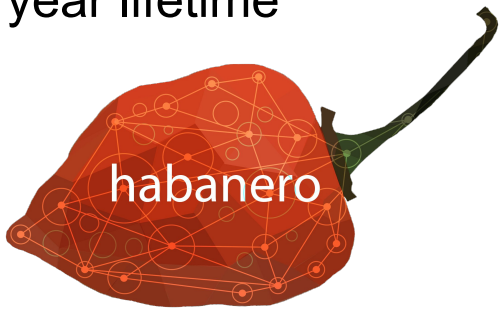
Kyle Mandli, *Chair, Research Computing Operating Committee*
Cesar Arias, *Manager, High Performance Computing, CUIT*

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

- Launched in 2016, expanded in 2017 with a 4 year lifetime
- Phase 1 was retired at the end of May 2022
- Phase 1 consisted of 222 nodes
  - 176 Standard
  - 32 High Memory
  - 14 GPU
- Habanero hardware is no longer under warranty and nodes will be disposed of as they die. We already have a number of nodes out of service.

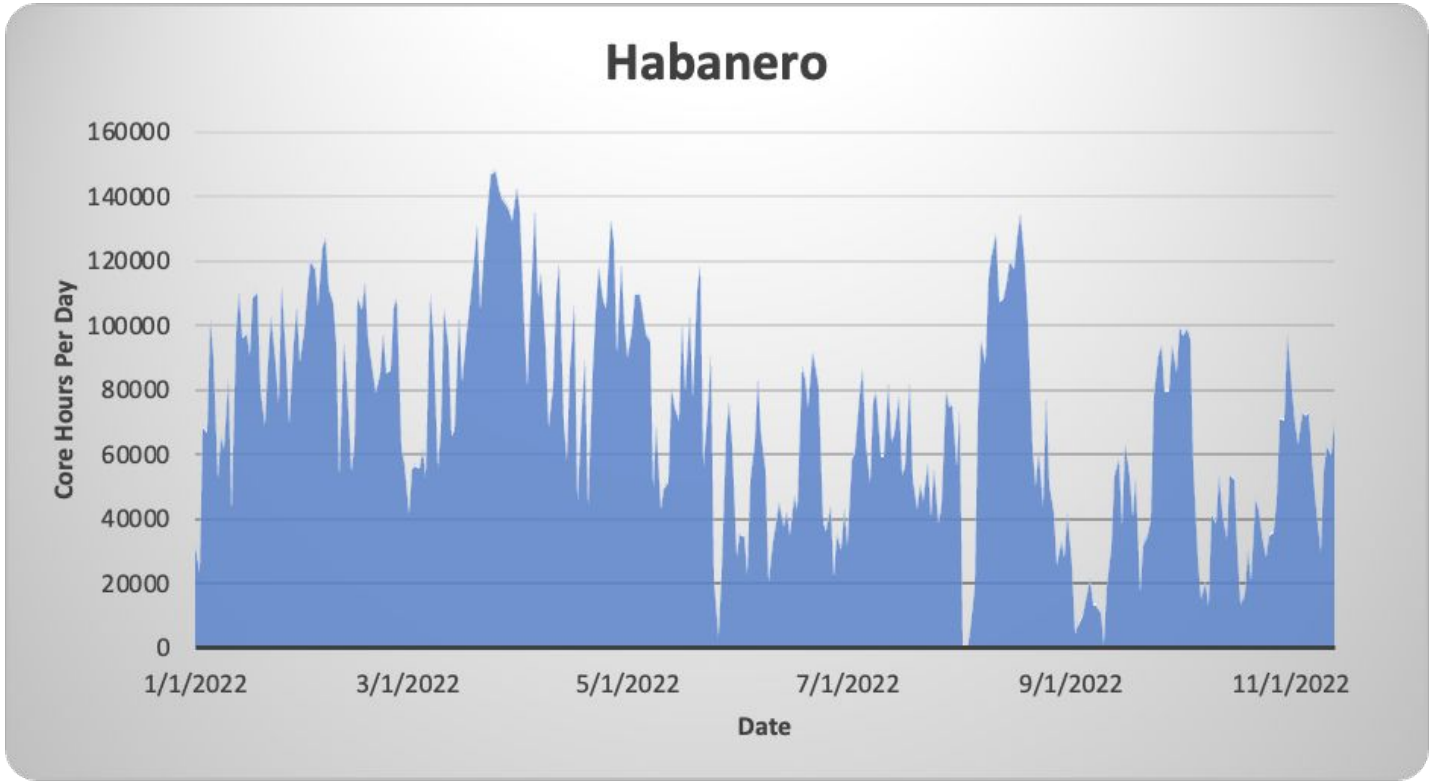


Habanero Retirement

- We will relocate the remainder of Habanero within the Data Center to make room for new equipment by end of 2022.

- Phase 2 will retire in April of 2023.

Habanero - Cluster Usage in Core Hours



Max Core Hours Per Day = ***148,698***

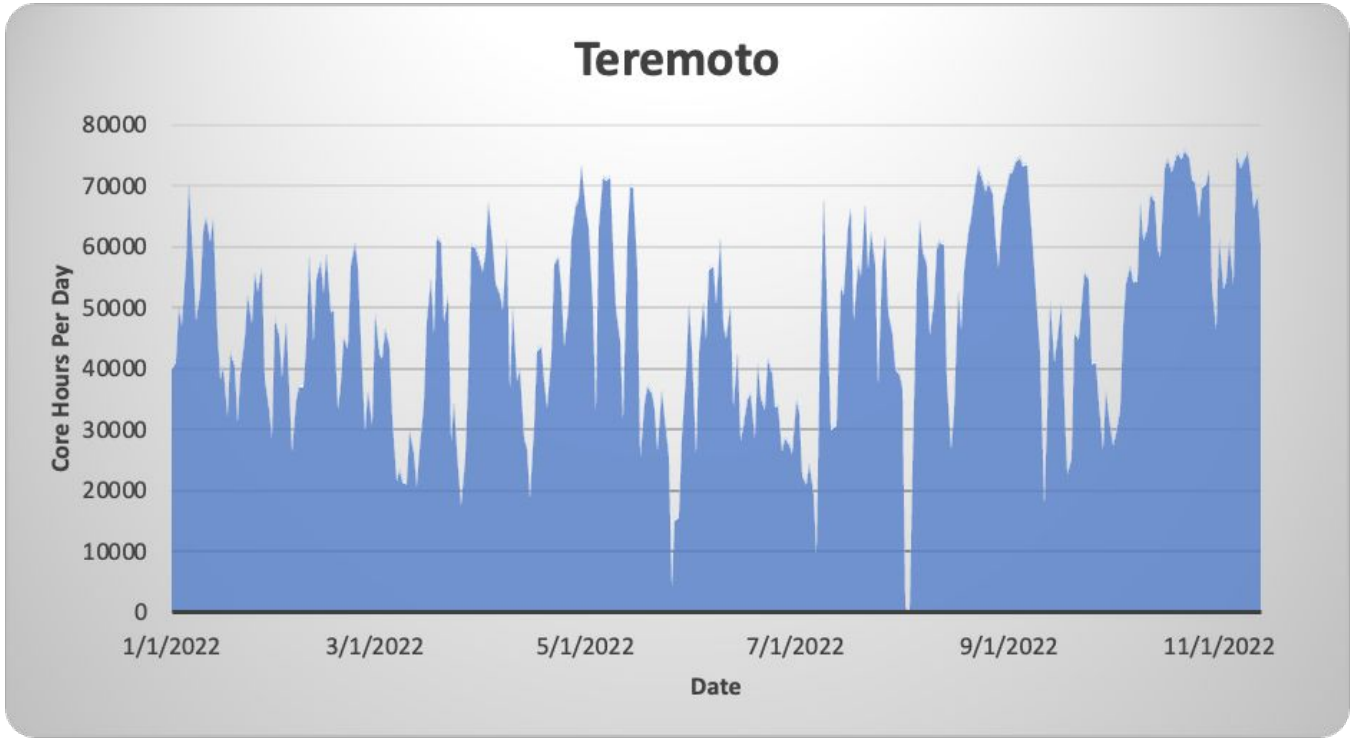
Total core hours in the past 12 months = ***23 million***

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



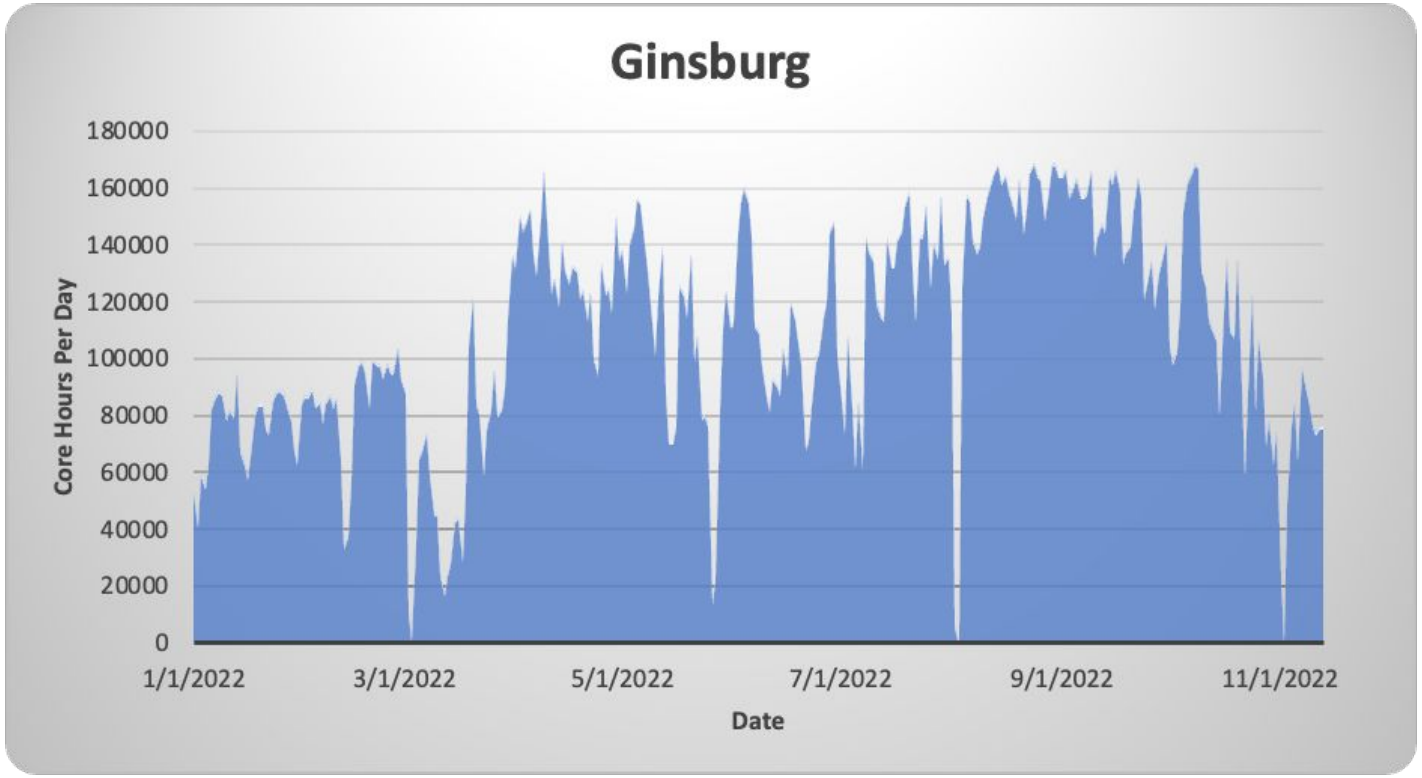
Max Core Hours Per Day = ***76156***

Total core hours in the past 12 months = ***15 million***

Ginsburg

- Ginsburg went live in February of 2021 with a $1.4 million joint purchase by 33 research groups and departments, consisting of 139 nodes with a total of 4448 cores.
- Ginsburg Expansion 1 (Phase 2) went live in March 2022 with over a million-dollars, expanding the cluster by 99 nodes with an additional 3168 cores.
- Ginsburg Expansion 2 (Phase 3) we anticipate to be available by end of year 2022 with a purchase total over six hundred thousand dollars expanding the cluster by 48 nodes with an additional 1536 cores bringing the cluster total to **286 nodes, 9152 cores and 39** GPU hardware accelerated systems.

Ginsburg - Cluster Usage in Core Hours



Max Core Hours Per Day = ***169,051***

Total core hours the past 12 months = ***34 million***

Shared High Performance Computing

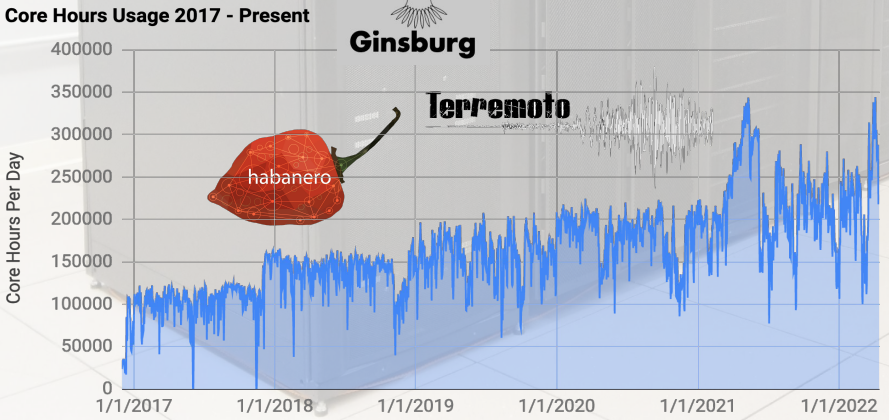
Providing Shared Compute
Since 2012

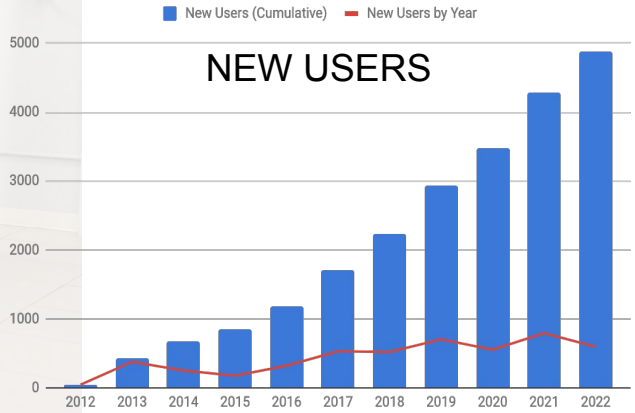
***Since 2012,** more than*
- 18 Million jobs run
- 363 Million core hours of compute provided
- 350 Peer-reviewed publications

***Currently** more than*
- 677 Compute Nodes
- 19,272 Cores
- 1236 TFlops
- 2.1 Petabytes of Storage

More than
70 Groups and Departments

Faculty-led Governance





Introductory training offered

COLUMBIA UNIVERSITY
INFORMATION TECHNOLOGY

GPU Cluster - Manitou

- Vendor is currently building and configuring mid-range GPU cluster. Supply chain issues have pushed delivery date to January 2023

- 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

RFP Committee

- Asking for volunteers to be a part of the RFP Committee for the next purchase
  - Time commitment for faculty: 4 meetings over 2-3 months

- RFP Committee timeframe is Feb/Mar 2023

- New storage solutions/vendors will be an area of focus.

HPC HIPAA Compliant Cluster?

- Preliminary investigation of HIPAA compliance
  - Best practices
  - Legal aspects
  - Separation of HIPAA workloads from other less secure dependent workloads
  - New cluster only
  - Internal resources
  - Adhere to physical security guidelines related to HIPAA
  - Separate storage solution with data encryption
  - Scrubbing and reprovisioning of nodes after job completion

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 Overview**

- **Research Computing Services Update**
  - Embedded Research Computing Support
  - Secure Data Enclave on-prem vs cloud
  - Globus - Connector Open Access
  - LabArchives Forum 2022
  - Scientific Software

Research Computing Services Overview

Embedded Research Computing Support
We provide embedded research computing support to CPRC, SSW, DSI, Stats, 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
This service helps organize and store laboratory data, provides information sharing, and enables 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.

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
A molecular biology software that allows users to plan, visualize, and document molecular biology procedures. Beginning last year, CUIT has offered our researchers the opportunity to purchase an annual SnapGene license at a **reduced price** through the University's multi-seat standard license.

Research Computing Services Update

- **Embedded Research Computing Support**
  - Provide on-site research computing support at Columbia research departments or centers
  - Hired and trained by CUIT
  - Provide curated services
  - Three new staff added since the beginning of this year
    - DSI, STATS, Psych Department

Research Computing Services Update

- **Secure Data Enclave on prem vs cloud**
  - **SDE on prem**
    - Since 2018 provides researchers with a Virtual cold room to analyze and collaborate on projects with restricted PII and PHI data sets
    - SDE demand is growing
      - The current environment was upgraded with the latest blades with the maximum RAM
      - Adding more blades and storage to address the need of high compute power and high storage capacity
  - **SDE in the cloud**
    - Not enough use cases for the cloud at this time
    - Data providers do not allow data in the cloud
    - Cost is high

Research Computing Services Update

- **Globus**
  - Current
    - Globus Standard Subscription
    - Globus Google Drive connector
  - Connector Open Access - allows unlimited use of all of the connectors
    - Amazon S3
    - Google Cloud Storage
    - HDFS
    - S3-compatible systems
    - (Caringo, Wasabi)
    - iRODS
    - Microsoft Azure Blob
    - Google Drive
    - Box
    - OneDrive
    - Spectra BlackPearl
    - Quantum ActiveScale
    - Ceph Object Stores
    - POSIX Archives

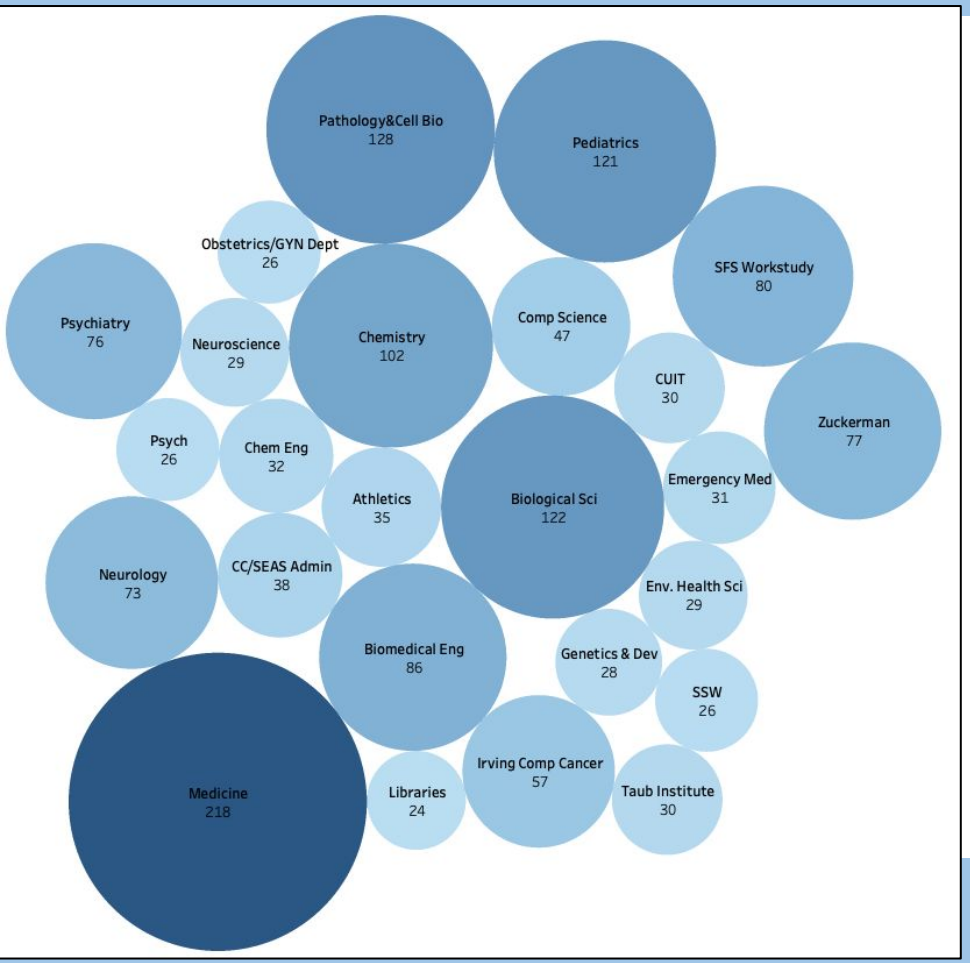
Research Computing Services Update

- **LabArchives Forum 2022**
  - Met experts from LabArchives in-person
  - Both campuses
    - CUIMC
    - Morningside
  - Two sessions
    - Introduction to LabArchives and User Case Studies
    - Advanced Topics with LabArchives

LabArchives Integrations

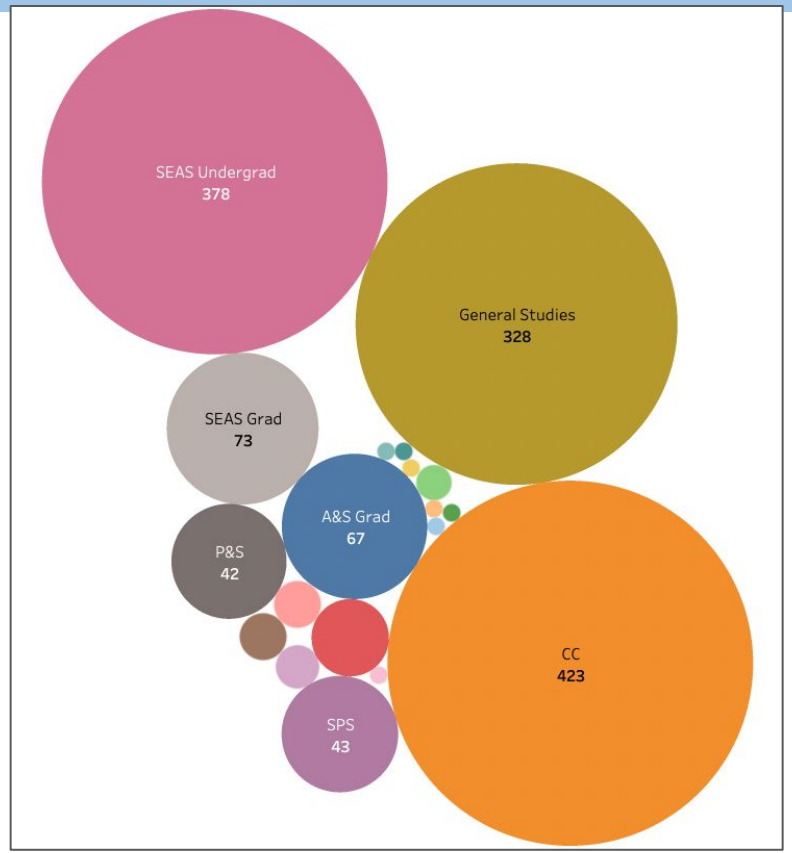
LabArchives integrates with many popular endpoint solutions.

Top 25 LabArchives Departments



- Most departments are based at CUIMC
- Several administrative units are also leveraging the tool (EVPR, Athletics, SFS Work Study)
- EVPR Handbooks will be accessible through LabArchives

Top 25 LabArchives Schools



- Most student users are based on Morningside
- Most student users are undergraduates

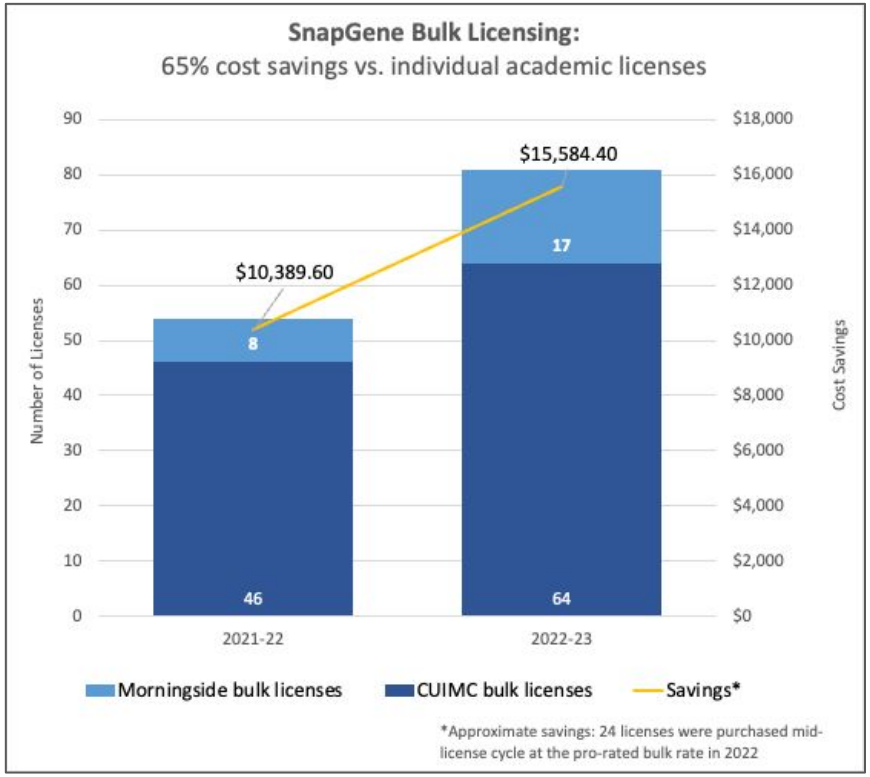
Columbia student groups

with the most LabArchives users

Scientific Software

- RCS Embedded Engineers scientific software survey in Spring 2022
  - Increased the number of users of SnapGene by 15%
  - Discovered desire for GraphPad Prism bulk license discount
    - a total of (176) group licenses under (@columbia.edu) with a combined total of (891) activations between these licenses
    - a total of 434 personal subscriptions
- Overleaf LaTex Editor
  - Working on an enterprise license, collaborating with CUIT and the Libraries
  - 9,017 confirmed Overleaf users registered with their institutional email address
  - Columbia's users often come from Computer Science, Physics, Economics, Electrical Engineering, and Mathematics

SnapGene bulk licensing

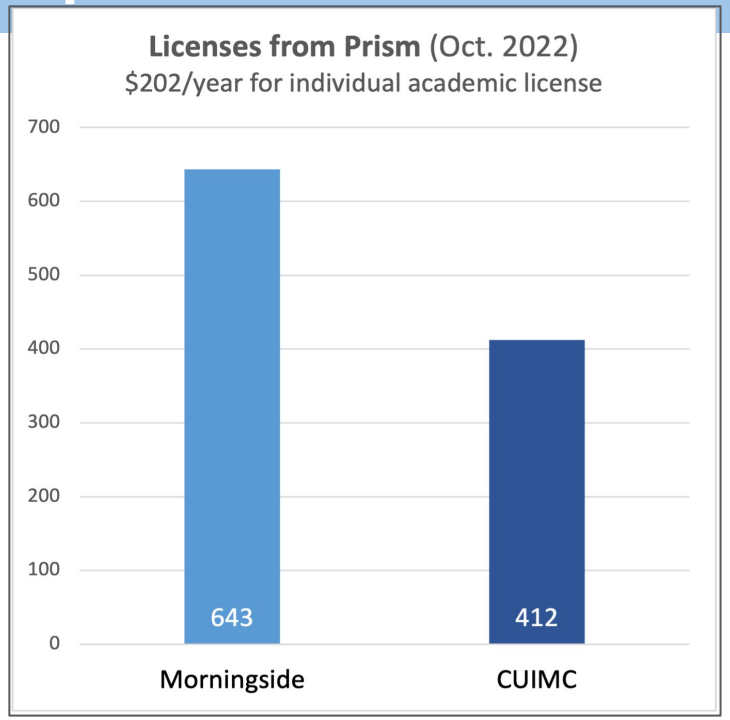


Top Users

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

SnapGene®
Software for molecular biology

GraphPad Prism bulk license opportunity



A bulk license for even 50 users represents nearly **50% savings** from the base individual academic license cost.

- Considering similar bulk license service for GraphPad Prism, a scientific graphing software that has 1K+ users across CUIMC and Morningside campuses.
- A DSI survey in Spring 2022 indicated over 250 users would be interested in a bulk Prism license

Prism 8
Analyze, graph and present your scientific work.

Research Computing Services

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

SRCPAC Fall 2022 Update

December 2nd, 2022

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.

Foundations Primary Activities

- **Novice boot camps**: 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

Foundations response to Covid

The challenges presented by COVID-19 required changes to the format of Foundations instruction, including the bootcamps. The shift to remote teaching provided some positive outcomes, including:

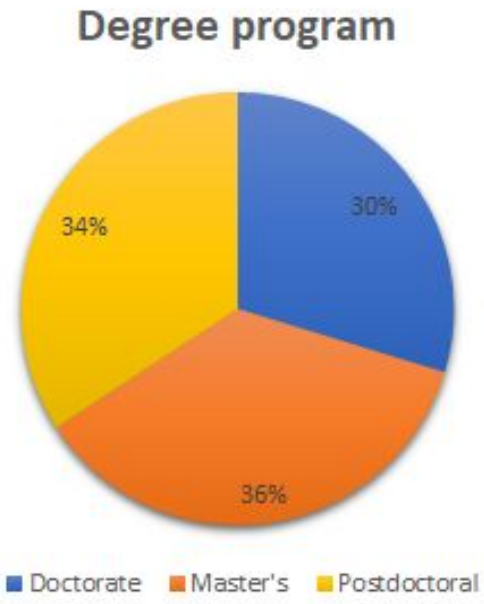
- Zoom format allowed helpers to address questions quickly in bootcamps
- Zoom format allowed for recordings & sharing with participants for all programming
- Able to support same number of workshop participants with a smaller number of workshop helpers
- Remote allows for greater participation in Python User Group
- Able to identify more specific needs for training by the way that researchers attended bootcamps

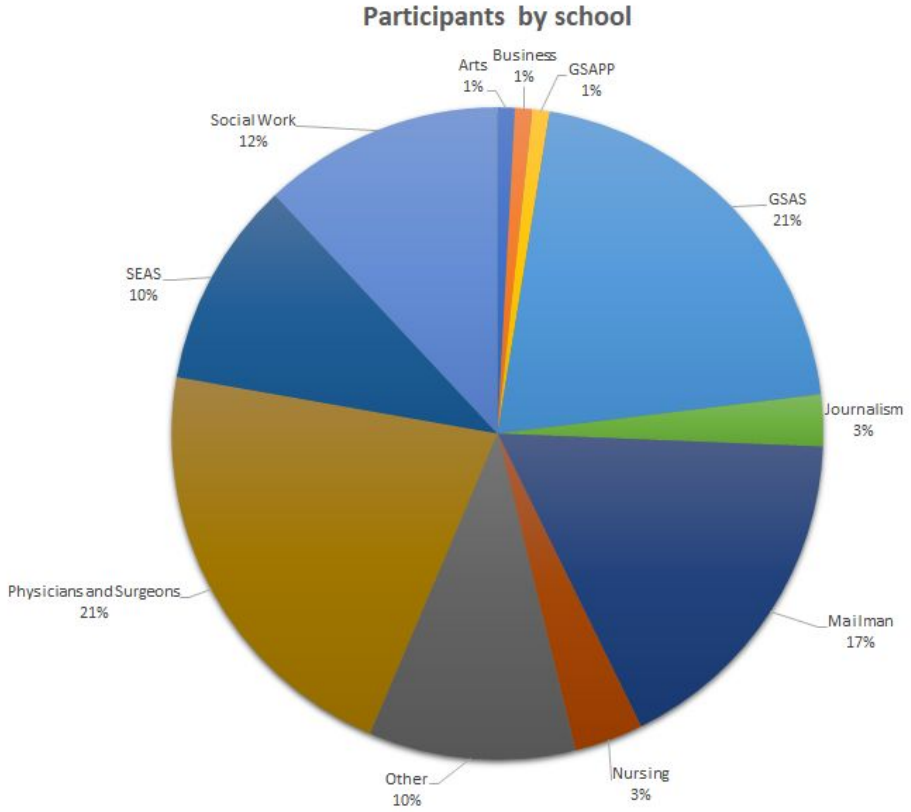
However there are some downsides to the transition as well:

- Software Carpentry is focused on in-person pedagogy
- Zoom fatigue
- Fewer (volunteer) instructors are comfortable with teaching in the zoom environment

Fall 2022 Bootcamp

Fall 2022 bootcamp offered remotely, 423 applications, trained 117 researchers





Next Bootcamp, Spring 2023

Spring 2023 bootcamp

- Scheduled for January 12th & 13th
- Return to in-person, held in Uris classrooms

Transitioned Python User Group to Data Club

What is Data Club?

- Offers a series of strategic lessons and a supportive meeting space for members of the Columbia community of all levels to learn, share, and implement the philosophies, tools, and methodologies for a data-driven project of their choosing.

- In bi-monthly meetings, provides a short instructive session, free-work time, and a collaborative and exploratory environment for researchers to pursue a personal project with the assistance of Columbia University librarians and their data-minded peers.

- Intentional curriculum designed for a mix of experience levels, building on skills acquired in bootcamps as well as curriculum for intermediate learners

Data Club Sessions

Fall Sessions:

- Introduction to Xarray (Python): Using labeled, multi-dimensional arrays in Python
- Introduction to HoloViz (part 1) (Python): Make more powerful, and more accurate visualizations
- Introduction to HoloViz (part 2) (Python)
- Introduction to Relational Databases (Python): Connect to SQL servers to query & update databases
- Exploratory Data Analysis with Observable (JavaScript): See your data
- Introduction to PyMC (Python): Build Bayesian models & do probabilistic programming in Python

Fall Foundations Partner Workshops

CUIT workshops:

- Introduction to Scripting
- Introduction to High Performance Computing

Library workshops:

- R workshop series: Basics; Data Structures; Data Visualization; Data Manipulation
- Research Data Management Plans & More
- Gitsertation: Scholarly project workflows with Git, Zotero, & Pandoc
- GIS in RStudio

Librarian Position

- Recruitment for a Foundations Coordinator was unsuccessful for a year and 4 months
- To increase the strength of the recruitment pools converted coordinator role to Librarian position
- Position is posted and are moving into active recruitment phase
- Benefits to new approach include:
  - Nationwide search process
  - Broader search pool
  - Greater recruitment incentives
  - Greater agency and responsibility in developing programming to working with partners

Librarian Position

Responsibilities

- 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.

Foundations for Research Computing: Assessment

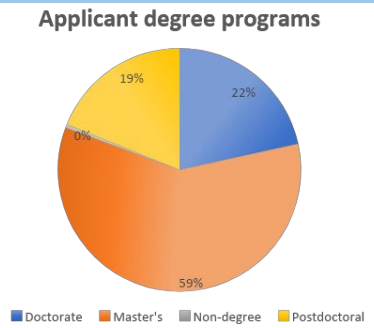
Demand has always been higher than what Foundations can provide. Understanding of why researchers apply will help identify approaches and new strategies to meet those needs outside the bootcamps.

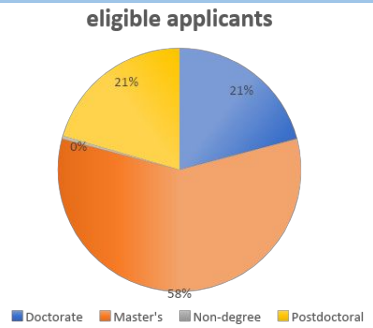
Assessment of Foundations boot camp data collected since 2018, with

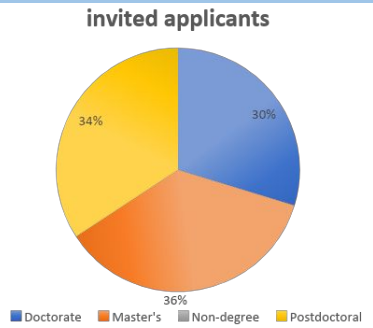
- Quantitative (i.e. school; dept; skill proficiency; degree program)
- Qualitative data (i.e. written reason for applying to bootcamp)

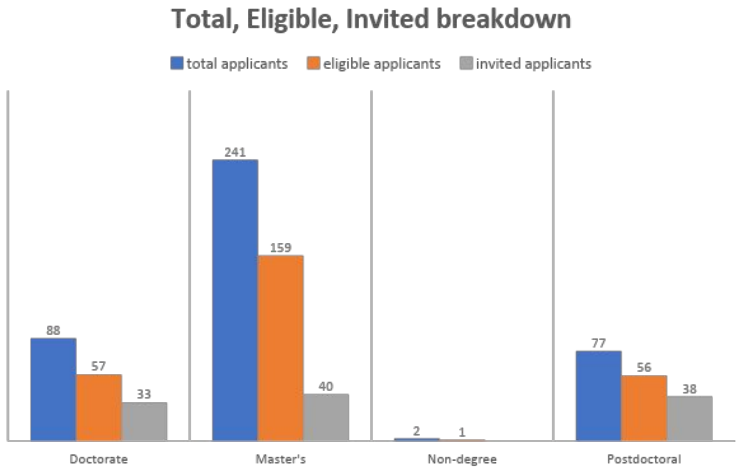
Foundations for Research Computing: Assessment

- Applicants answer four questions about familiarity of Git, Unix, Python, and R. Rated on a scale of 1-5
- The scores are averaged, and participants with an average of 2.5 or less are eligible for the training
- Invitation weighted toward expressed interest or need in research

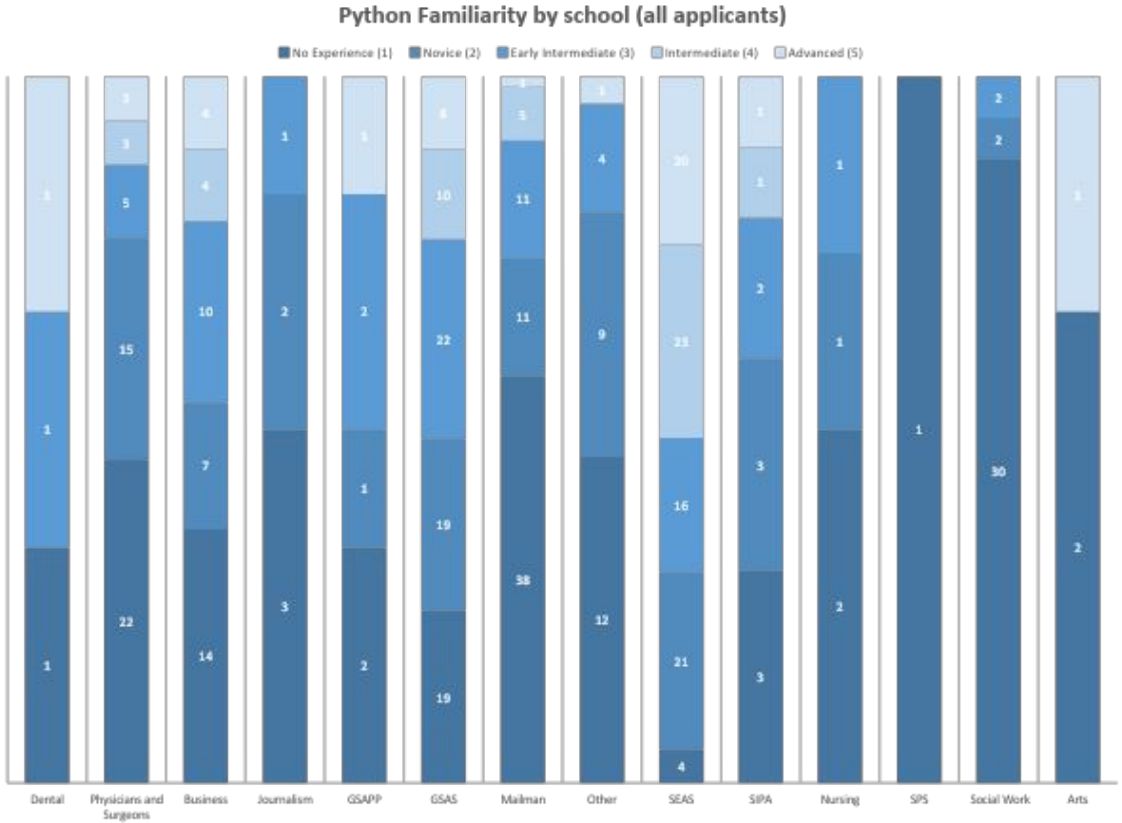








Foundations for Research Computing: Assessment



Future goals for the program

- Recruit the Foundations Program Librarian
- Continuation/expansion of current programs
  - (bootcamps, Data Club, workshops…)
- Develop a more robust assessment plan to address demand/scaling issue
- Seeking input/new partners in guiding future directions of Foundations
- Happy to take any questions