# 2024 Annual Meeting
# Research Computing Executive Committee (RCEC)

Friday, May 10

**Alexander Urban,** *Chair - Shared Research Computing Policy Advisory Committee (SRCPAC)*
**Jeannette Wing,** *Chair of RCEC, Executive Vice President for Research*
**Marc Spiegelman,** *Chair - Foundations for Research Computing Advisory Committee (Foundations)*
**Hod Lipson,** *Co-Chair of RCFC*
**Darcy Peterka,** *Co-Chair of RCFC*

# RCEC Agenda

**Welcome & Introductions**

- Alexander Urban, Chair of SRCPAC, *Assistant Professor of Chemical Engineering*

**Empire AI**

- Jeannette Wing, Chair of RCEC, *Executive Vice President for Research*

**Shared High-Performance Computing Update**

- Alexander Urban, Chair of SRCPAC

**Foundations for Research Computing Update**

- Marc Spiegelman, Chair of the Foundations for Research Computing Advisory Committee; *Professor of Earth and Environmental Sciences and Professor and Chair of Applied Physics and Applied Mathematics*

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

- Hod Lipson, Co-Chair of RCFC, *Professor of Mechanical Engineering*
- Darcy Peterka, Co-Chair of RCFC, *Zuckerman Mind Brain Behavior Institute*

# Empire AI

**Jeannette Wing**

*Chair of RCEC, Executive Vice President for Research*

# SRCPAC – Origin and Charter

*Excerpt from the SRCPAC Charter, November 9, 2011:*

"The Shared Research Computing Policy Advisory Committee (SRCPAC) will be a **faculty-dominated group focused on a variety of policy issues related to shared research computing on the Morningside campus**. As the use of computational tools spreads to more disciplines to create, collaborate, and disseminate knowledge, there is a commensurate rise in the costs of establishing and maintaining these resources. Shared resources have proven to leverage those available to individuals or small groups, but require careful consideration of the policies governing the shared resource and the basis of the operating model.

While final authority and responsibility for such policies customarily rests with the senior administrators of the University, it is vital that the **research faculty examine and recommend the policies and practices they deem best suited to accomplishing the research objectives**."

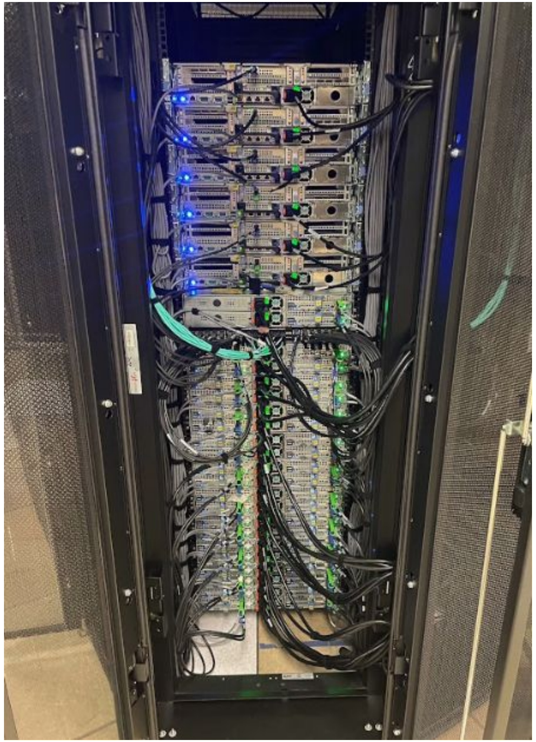
# High Performance Computing Updates

**Alex Urban**

*Chair, Shared Research Computing Policy Advisory Committee (SRCPAC)*

# HPC in Summary

- Over 2200 faculty, students, and researchers used the Shared HPC service of nearly 16,000 cores this past year.
- Utilization and adoption continue to grow since its founding in 2012.
- Opportunity exists to create a more robust free and educational tier, but would require additional personnel.



# Current Model

- The **University** provides facilities, cooling, and electricity.
- **CUIT** provides
  - High density racks, rack support
  - Data center staff
  - HPC system administration, engineering, and support staff
- **Researchers** pay for hardware, storage, software, networking, cables.
- Hardware has a limited **lifetime** decided at purchase time and tied to maintenance contracts (~5 years)

# Shared High Performance Computing

*Providing Shared Compute*

Since 2012

Faculty-led Governance

***Currently** more than*

- 570 Compute Nodes
- 15,968 Cores
- 518 TFlops
- 2.3 Petabytes of Storage

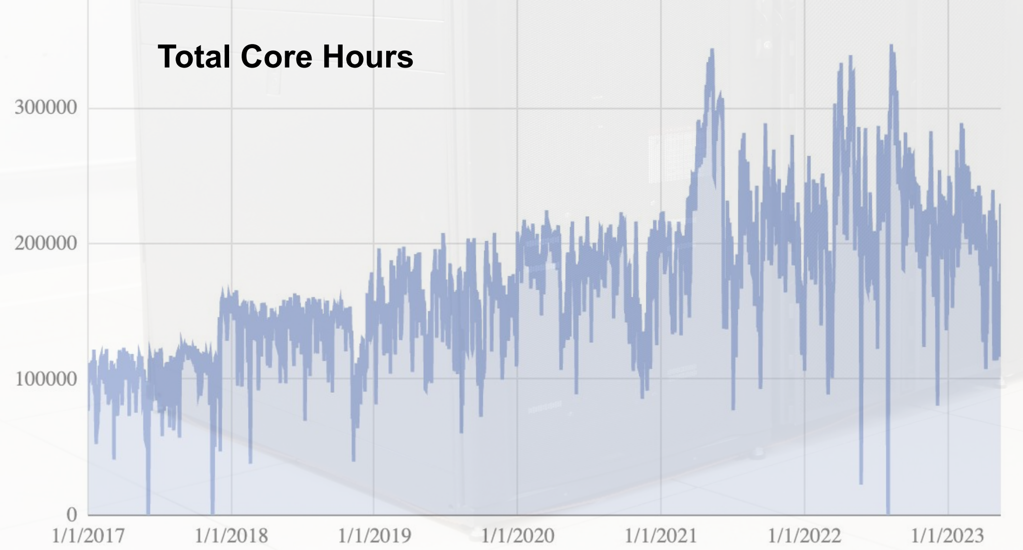
***More than***

- 20 Million jobs run
- 400 Million core hours of compute provided

***More than***

- 180 Group and Department shares
- 4900 users since 2017

## Introductory training offered



**Since 2017**

**Edu Tier**

Total Users: 530 students

Total Use: 1,744,956 core hours

**Free Tier**

Total Users: 305

Total Use: 1,846,479 core hours

Columbia University
Information Technology

# Current HPC Footprint

## Terremoto Phase 2

- 18 Standard Nodes (192 GB)
- 4 High Memory Nodes (768 GB)
- 1 GPU 1x V100
- 3 GPU 2x V100

## Ginsburg Phases 1, 2, and 3

Ginsburg has 286 nodes with a total of 9,152 cores (32 cores per node)

- 191 Standard Nodes (192 GB)
- 56 High Memory Nodes (768 GB)
- 18 GPU 2x RTX 8000 GPU modules
- 4 GPU 2x V100S GPU modules
- 9 GPU 2x A40 GPU modules
- 8 GPU 2x A100 GPU modules

## Manitou - GPU Cluster

The cluster has 15 GPU nodes:

- 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

## Insomnia

Insomnia has 40 nodes with a total of 3,200 cores (80 cores per node)

- 24 Standard Nodes (192 GB)
- 10 High Memory Nodes (768 GB)
- 3 GPU 2 x L40
- 2 GPU 1 x H100 (backorder)
- 1 GPU 2 x H100 (backorder)

## Free Tier

A portion of retired hardware, on a best-effort basis

# Who is buying in?

### Terremoto Phase 1
Chemical Engineering
Mechanical Engineering
Computer Science
APAM
Civil Engineering
Statistics
Astronomy

### Terremoto Phase 2
Social Science Computing Consortium
Irving Institute for Cancer Dynamics
Statistics
Computer Science
Lamont-Doherty
Chemical Engineering
Zuckerman Institute
Department of Medicine
Chemistry

### Ginsburg Phase 1
Ocean Climate Physics
Earth and Environmental Sciences
Mechanical Engineering
APAM
Biomedical Engineering
Chemical Engineering
Electrical Engineering
Astronomy
Biological Sciences
Chemistry
Psychology
Psychiatry
Neuroscience
Irving Institute for Cancer Dynamics
Computational Electrochemistry

### Ginsburg Phase 2
Biological Sciences
Statistics
Astronomy
LDEO
Ecology, Evolution, and Environmental Biology
Biomedical Engineering
CCCE
Irving Institute for Cancer Dynamics
Physics
Astrophysics
Computer Science

### Ginsburg Phase 3
Astrophysics
Earth and Environmental Engineering
Irving Institute for Cancer Dynamics
SSCC
APAM
Natural Sciences
SEAS Dean's Office
Zuckerman Institute
Chemical Engineering
Biostatistics
Environmental Health Sciences
HICCC

### Insomnia
MSPH IT
Physics
Industrial Engineering and Operations Research
Irving Institute for Cancer Dynamics
Earth and Environmental Engineering
Statistics
Chemical Engineering
SIPA Center on Global Energy Policy
Biostatistics
Computer Science
Biomedical Engineering
APAM
Ecology, Evolution and Environmental Biology
Biological Sciences
Astrophysics

### Manitou
Systems Biology
Computer Science

# High Performance Computing Capacity

- FOUR factors affect High Performance Computing Capacity

*Space* *Cooling* *Power* *Personnel*

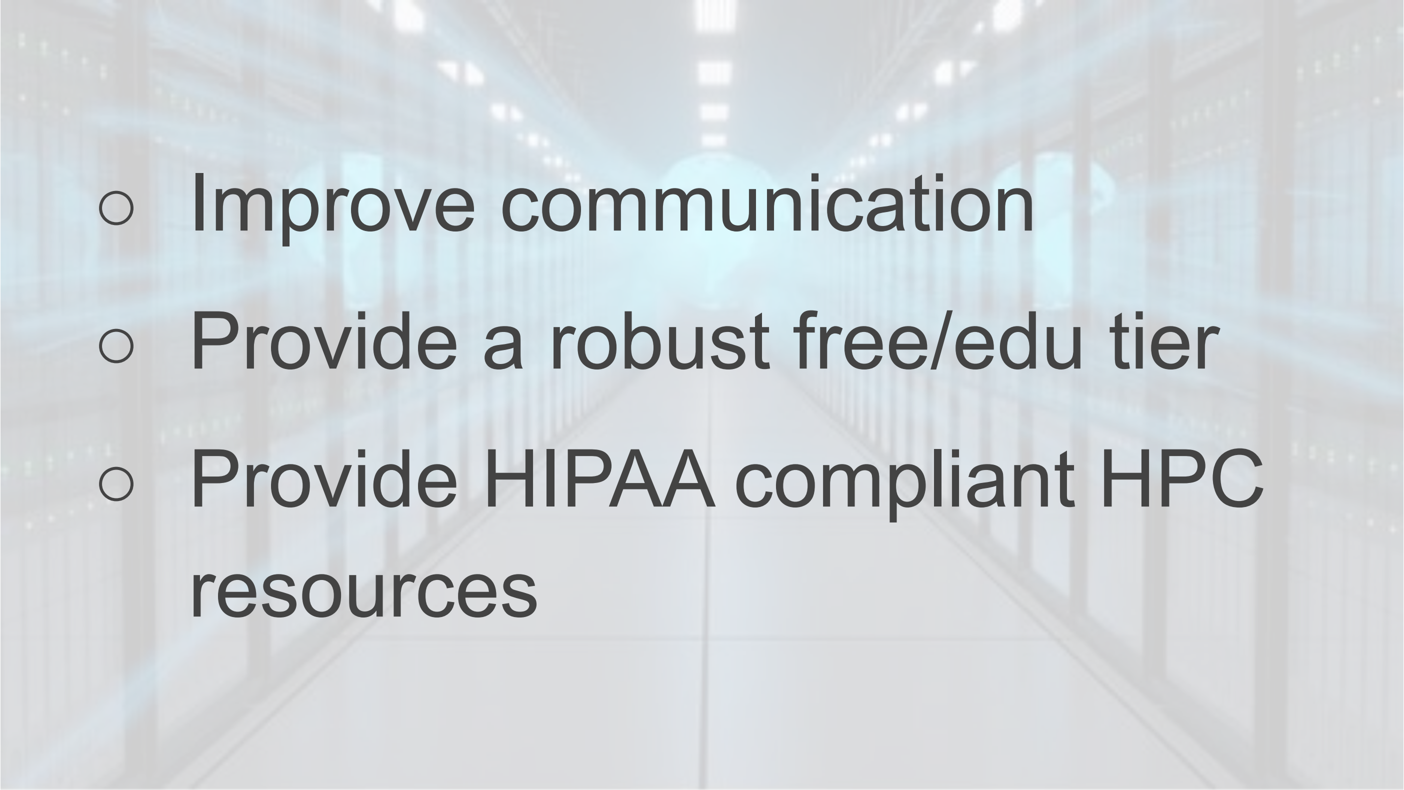
- Space: We are currently occupying 13 of the 16 HD racks. Retiring hardware keeps racks rotating. Extending the life of nodes past 5 years would push against capacity
- Power:
  - Capacity: 16 HD racks fully loaded at 25kW = 400kW
  - We are currently using approximately 250kW
- Cooling: Expanding chilled water beyond the existing 16 racks will require capital investment.
- Personnel: Expanding training and or extending the life of nodes might require more staff.

# Insomnia - new purchasing model

NEW PLANS FOR CLUSTER MANAGEMENT AND PURCHASING

- Moving to a single cluster with more flexible ability to join rather than a new cluster every couple of years.
  - One storage system
  - Better rack utilization
  - Central provisioning
- Working to shift to ordering quarterly with an ability to purchase at a set price from the current annual lengthy purchase rounds.
- Offering a 1/4 share of standard node to address high prices.
- Rental option still available.
- Communication to come out soon!

# Areas for Expansion/Improvement



- Improve communication
- Provide a robust free/edu tier
- Provide HIPAA compliant HPC resources

# We are Overhauling the SRCPAC Website

- Feedback tells us that many colleagues on campus are unaware of existing shared research computing resources
→ see long-term strategy at the end of this meeting
- Currently restructuring SRCPAC website to provide clearer overview, especially of the HPC tiers for new users
- The Research Computing Services (RCS) have been compiling online training materials for intermediate to advanced users (https://www.cuit.columbia.edu/about-research-computing-services)

# A Robust Free/Edu Tier Would Require Minimal Resources

Presently our HPC resources lack a robust free/edu tier. We offer access to retired equipment with minimal maintenance or user support.

- Originally, the free tier consisted of four computer nodes jointly purchased by Engineering and Arts & Sciences.
- Support for aging, out-of warranty equipment requires more monitoring, administrative support, and more hands-on support.
- **Additional personnel** would enable us to provide prompt support and guidance to new users.
- Or/and **new hardware** (1-2 nodes) with maintenance contracts would simplify maintenance.

# A HIPAA Compliant HPC Cluster will Require Space and Staff

Given the increasing importance of data security and compliance, there is demand for HPC resources capable of handling sensitive data.

- Looking for a **location** to house a HIPAA-compliance cluster
- Building/maintaining a HIPAA compliant HPC Cluster will require additional **HPC staff** with expertise in security and compliance.

HIPAA COMPLIANT

# Foundations Update

**Marc Spiegelman,** Chair of the Foundations for Research Computing Advisory Committee

# Foundations Mission

**Foundations for Research Computing** provides an **informal introduction** for Columbia University graduate students and postdoctoral scholars to the 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 complement our investment in hardware, software and systems administration

# Initial Design of Foundations

- **Novice Level**
  - Institutional Partnership with Software Carpentry
  - SC Bootcamps
- **Intermediate Level**
  - RCS HPC tutorials
  - Intensives and Workshops
  - Python User Group/Python Club
  - Partnered with Departmental Training (e.g. MechE, CUIMC)
  - Other modes (Distinguished Lecture series, CIG)
- **Advanced level**
  - Coordination with departmental curriculum

# Initial Design of Foundations

- **Novice Level**
  - Institutional Partnership with Software Carpentry
  - SC Bootcamps
- **Intermediate Level**
  - RCS HPC tutorials
  - Intensives and Workshops
  - Python User Group/Python Club
  - Partnered with Departmental Training (e.g. MechE, CUIMC)
  - Other modes (Distinguished Lecture series, CIG)
- **Advanced level**
  - Coordination with departmental curriculum

# New Hires

**Anne Cong-Huyen**, Ph.D.
*Director of Digital Scholarship*

Her portfolio includes Research Data Services, Academic Commons, Library Publishing Services, the Digital Humanities Center, and related services.

*Anne was previously the Director of Digital Scholarship at the University of Michigan Library. She has a PhD from UC Santa Barbara.*

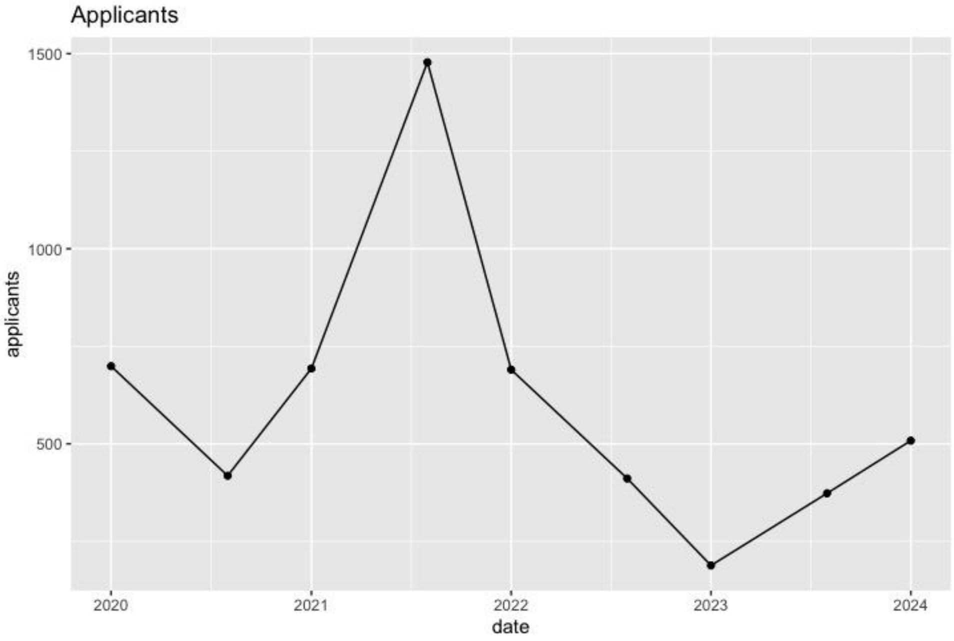
**Dan Woulfin,** Ph.D., M.L.S
*Computational Research Instruction Librarian*

Dan oversees the Library's instructional program around computational literacy and practical skills. He works with partners in CUIT, EVPR, DSI, and others across campus.

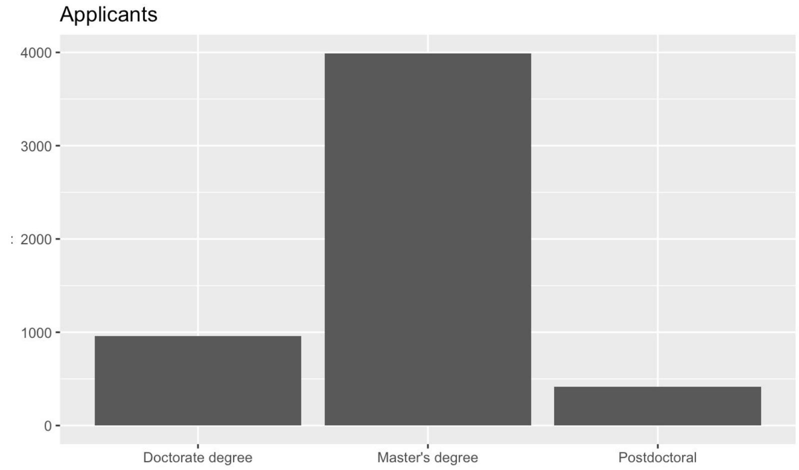
*Dan earned his Ph.D. from Stony Brook University - SUNY and his Masters of Library Science from Queens College - CUNY.*

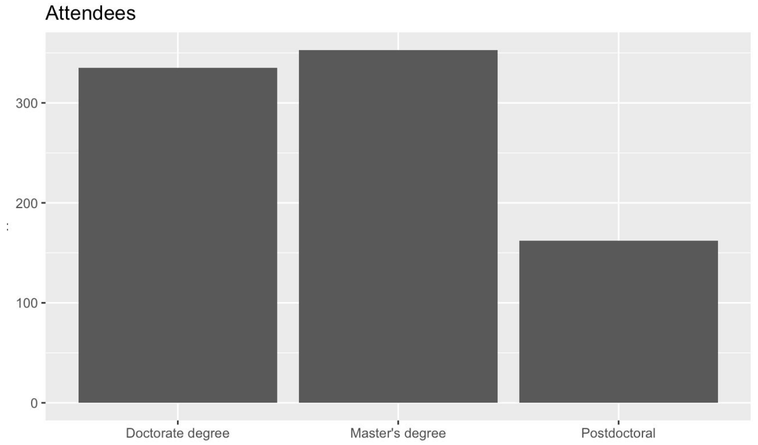
# Software Carpentries Workshops for Beginners

- 15 Workshops since Aug 2018 (2-3/year), initial demand ~800 applicants, ~120 attendees per bootcamp.
- 2020 shifts to online only for Covid
- Initial coordinator P. Smyth leaves mid 2021
- Interim leadership by the libraries
- Return to in-person, January 2023 (188 applicants, 28 attendees accepted)
- Dan and Anne hired (August/September 2023)
- By January 2024 demand returned (508 applicants, 75 attendees invited, 58 accepted)
- **Next workshop: August 20-21 2024**



# Foundations: Summary of Impact

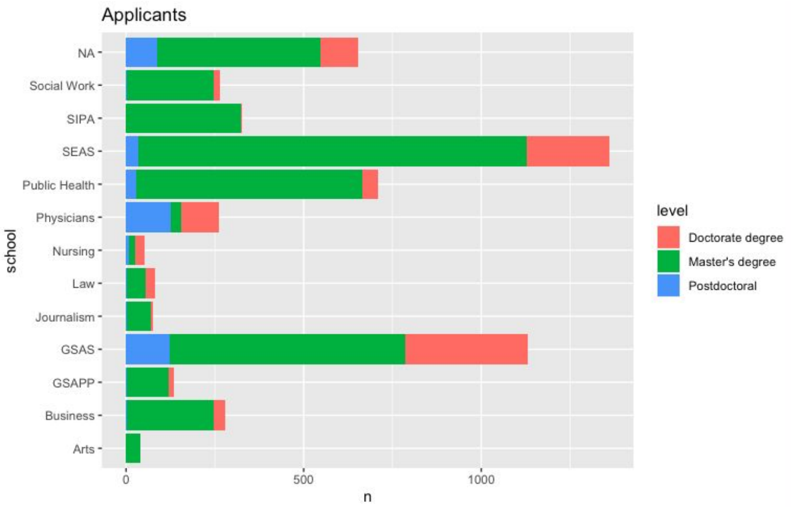


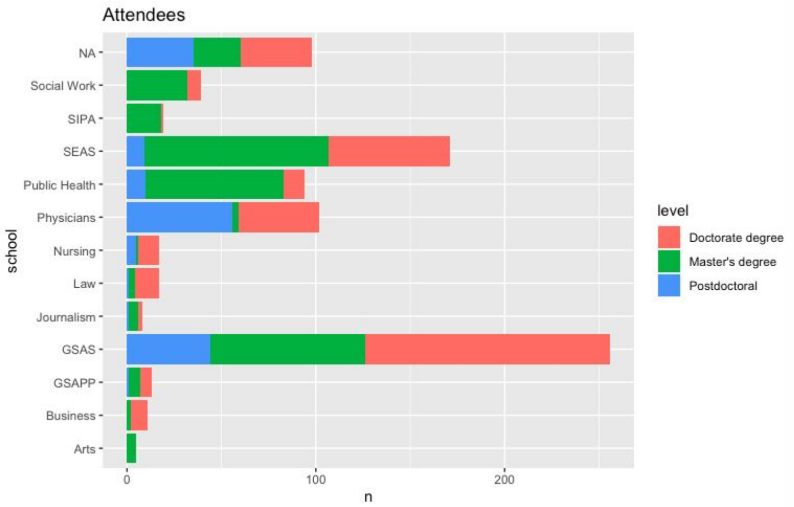


- **Who is the audience (applicants and attendees)?**
  - Since January 2020, we've had **5,458** applicants
    - Master's students - 3,991 applicants (73.4%) | Ph.D. Students (961 - 17.7%) | Postdocs (415 - 7.63%)
  - Attendees - **850** since January 2020
    - Master's students - 353 attendees (41.5%) | Ph.D. Students (335 - 39.4%) | Postdocs (162 - 19.1%)

# Foundations: Summary of Impact (By School)

| | GSAS | SEAS | SIPA | Business | Social Work | GSAPP | Law | Journalism | Arts | Nursing | Public Health | Physicians |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Applicants** | 1130 | 1362 | 325 | 280 | 264 | 135 | 81 | 75 | 40 | 52 | 709 | 261 |
| **Attendees** | 256 | 171 | 19 | 11 | 39 | 13 | 17 | 8 | 5 | 17 | 94 | 102 |

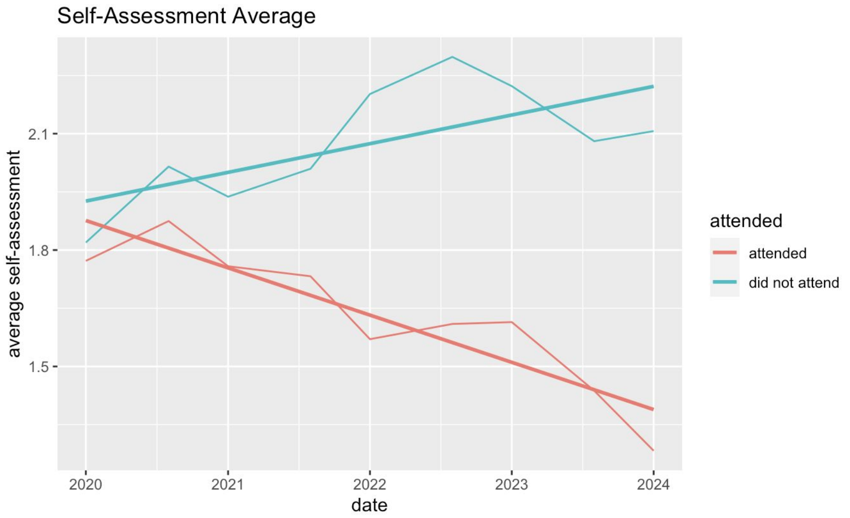




# Foundations: Summary of Impact

## How to scale to meet demand?

- The Carpentries model is very labor intensive and limits our capacity
  - Carpentries workshops requires a ratio of 1 instructor/helpers for every 8 learners.
  - The number of trained volunteers has decreased (55% attrition).
- Capacity decreased post-Covid to two tracks, Python and R, with 60 total learners maximum
- Based on self-assessment of applicants, the average technical ability has risen, so most applicants have some programming experience. However, the average attendee has little to no experience.



**We are piloting new workshops with the Carpentries to serve previously underserved populations (social sciences and intermediate learners)**

# Foundations: Issues for Moving forward

**Foundations was originally designed to provide a informal, tiered path for graduate students and postdocs to develop basic and intermediate computational skills to prepare them for advanced problem solving. We remain committed to this mission.**

**After assessing the past program and trends, however, the following issues *persist***

- Maintaining contact with faculty and adjusting to changing computational needs
- Scaling capacity to meet demand
    - Addressing the labor-intensive model and curriculum we currently use for novice users
    - Providing training for the increasing numbers of intermediate learners.

# Working Towards a New Framework

Foundations is working to develop a more sustainable, expanded Foundations program

Our current directions include the following steps:

- Re-engaging the Foundations faculty advisory group
- Centering Computational Literacy as a structuring focus in the Library
    a. Build capacity within the Library in computational methods/coding expertise
    b. Piloting and building new Library workshops around computation
    c. Curating additional resources and linking learners to new and existing learning opportunities
- Developing additional resources and opportunities for self-paced learning and/or more varied learning modalities
- Partnering with other computational efforts (e.g. Training efforts in Long-term strategy)

# 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 administration

# Foundations Primary Activities

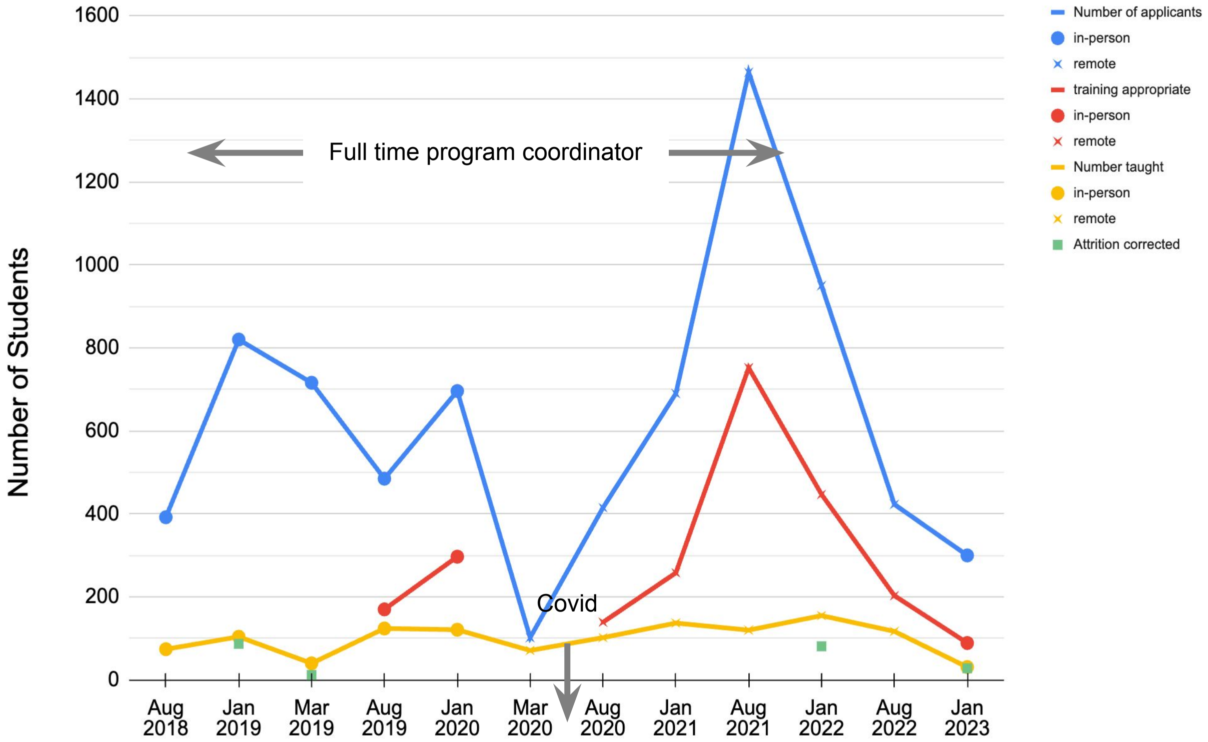
- **Novice trainings**: 2 day training based on Software Carpentry curriculum for novice learners, learning Git, UNIX, and either R or Python
- **Data Club**: revamping of Python Users Group: twice-monthly meeting for those using computation in their research or interest about specific, more advanced 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

Jan 2023 in-person training – first since Jan 2020

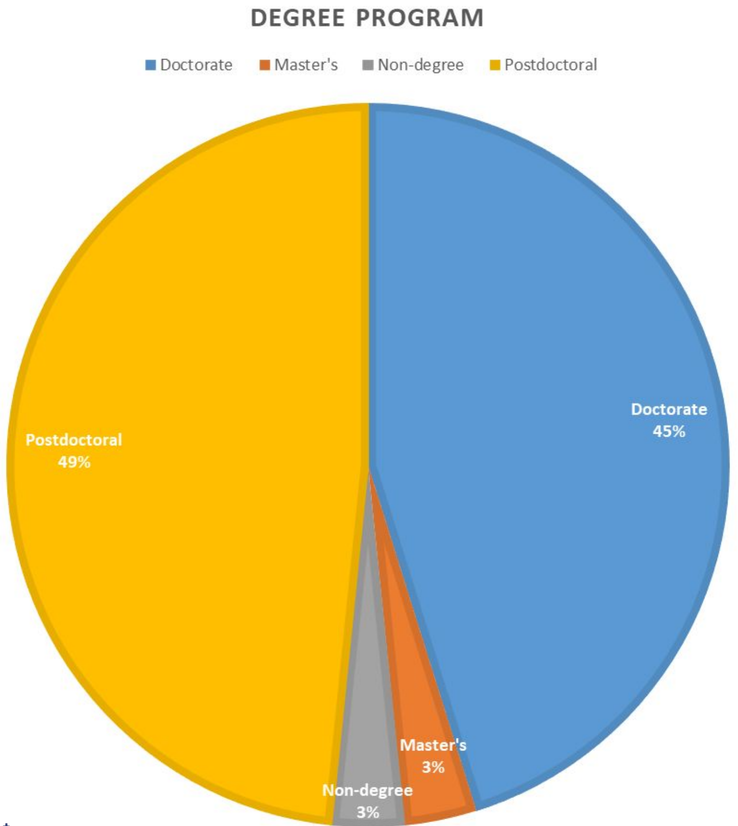
# Novice Training Data

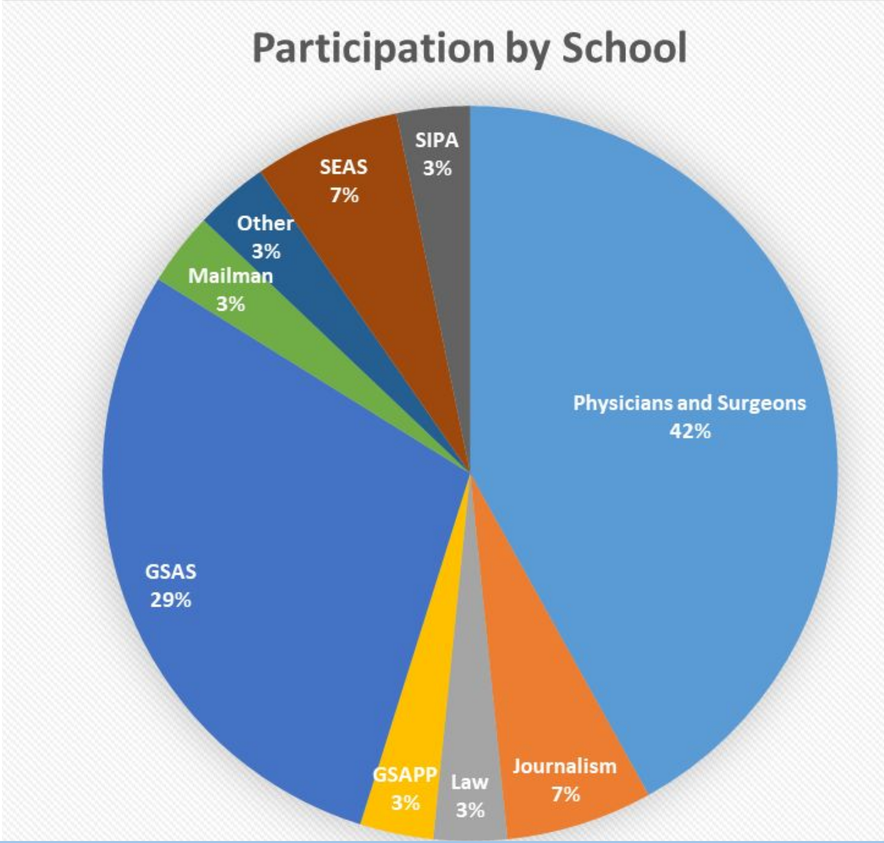


## Some Observations

- 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 Novice Training (31 participants)





# 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

# Thank you

# HPC 2023 Purchase Round - Pricing Menu

| | LAST YEAR | 2023 estimate |
|---|---:|---:|
| **Standard Server (512 GB)** | $7,404 | $14,105 |
| **High Memory Server (1 TB)** | $14,922 | $16,892 |
| **GPU server with 2 x A40** | $16,808 | $21,340 |
| **GPU server with 2 x A100** | $25,661 | $29,774 |

## Servers Feature

Dual Xeon Platinum 8640Y+ processors (2 GHz, 40 cores each, 80 cores per server), 512 GB Memory

This is a significant increase in cores and memory over last year’s model (80 cores vs 32 cores)

## Prices Include

- Infrastructure-related costs
- Networking
- Scheduling software
- 5-year support and maintenance