High Performance Computing Update

Dr. Chris Marianetti, *Chair of SRCPAC*
Associate Professor
Applied Physics and Applied Mathematics

High Performance Computing Update

Agenda

- Current HPC status
  - Habanero Phase I Retires June 1
  - Annual Opportunity to join latest HPC, aka Ginsburg
  - First-time Opportunity to join Manitou GPU-only Cluster
- NIH G20 Final Site Visit
- Future of HPC
- Chair of SRCPAC

Shared High Performance Computing

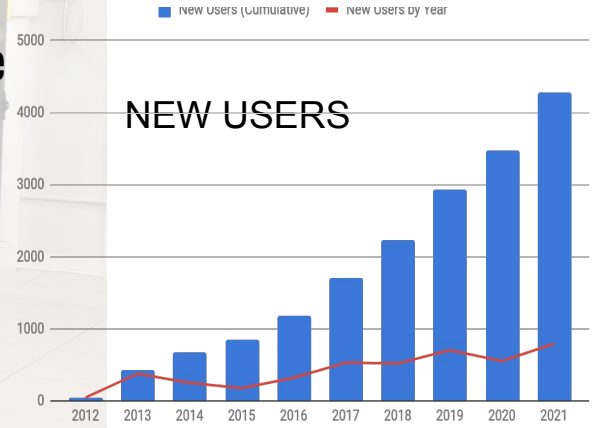
Providing Shared Compute
Since 2012

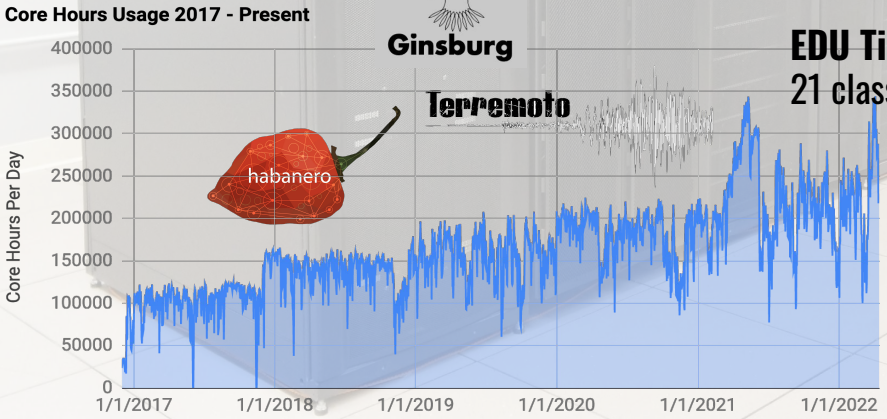
***Since 2012,** more than*
- 18 Million jobs run
- 314 Million core hours of compute provided
- 350 Peer-reviewed publications

***Currently** more than*
- 677 Compute Nodes
- 18,176 Cores
- 1236 TFlops
- 2.1 Petabytes of Storage

More than
70 Groups and Departments

Faculty-led Governance





EDU Tier
21 classes with 998 students

Introductory training offered

COLUMBIA UNIVERSITY
INFORMATION TECHNOLOGY

Current HPC Footprint

	Status	Nodes	Cores	Total $	Comments
Habanero Phase I	Retired June 1 2022	222	5328	$1.55M	Maintaining a portion as free/edu tier
Habanero Phase II	Active until Spring 2023	80	1920	$745K	
Terremoto Phase I	Active until Dec 2023	110	2640	$1.3M	
Terremoto Phase II	Active until Dec 2024	27	648	$318K	
Ginsburg Phase I	Active until Dec 2025	139	4448	$1.4M	
Ginsburg Phase II	Active until Dec 2026	99	3168	1.07M	
Ginsburg Phase III	Go live late 2022				Currently taking orders
Manitou (GPU)	Go live Fall 2022	13	1248		Currently taking orders

EDU Tier

- In 2016, A&S and SEAS invested $12K each in 2 standard nodes each to be allocated for the EDU tier (Habanero Phase I)
- This hardware will be retired Spring 2023
- 21 Classes since Habanero Launch with 998 students
- Also supports the biannual Intro to HPC Workshop series hosted by CUIT
  - Intro to Linux
  - Intro to Scripting
  - Intro to HPC

HPC 2022 Purchase Round - Schedule

- Buy-in period to join new cluster open through **June 15, 2022**
  - **Prices are significantly higher than last year due to supply chain issues**
- Purchase Order to be issued in July/August 2022
- Go-live of new equipment planned for late Fall 2022

NEW GPU CLUSTER, **Manitou**

- Includes option to join new, mid-grade, GPU cluster, named **Manitou**
- An anchor tenant provided funds to start up the new GPU cluster

HPC 2022 Purchase Round - Pricing Menu

	LAST YEAR	NOW
Standard Server (192 GB)	$7,850	$10,611
High Memory Server (768 GB)	$12,750	$20,591
GPU server with 2 x A40	$13,750	$19,987
GPU server with 2 x A100	$18,850	$32,042

Servers Feature

- Dual Cascade Lake Intel 6226R processors (2.9 GHz speed, 32 cores per server)
- EDR 100 GB/s Infiniband
- 480 GB SSD drives

Prices Include

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

HPC 2022 Purchase Round - Pricing Menu

Manitou GPU Cluster	
GPU Server with 2 x A6000 (1) EPYC 7543P 2.8GHz 32-Core, 256GB DDR4-3200 (8x 32GB) memory, 100GB/s Infiniband, (1) 512GB SATA SSD	$22,354
GPU Server with 4 x A6000 (1) EPYC 7543P 2.8GHz 32-Core, 256GB DDR4-3200 (8x 32GB) memory, 100GB/s Infiniband, (1) 512GB SATA SSD	$30,729
GPU Server with 8 x A6000 (2) EPYC 7643 2.3GHz 48-Core, 1TB DDR4-3200 memory, 200GB/s Infiniband, (2) 15TB U.2 Gen 4 NVMe SSD	$64,132

Prices Include

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

NIH G20 Final Site Visit

- NIH G20 Final Site Visit was today, 5/31/22

- Summary: $10M to renovate infrastructure serving an existing 5,000 square foot central data center to provide a new Core Research Computing Facility (CRCF) that will consolidate computational resources and improve data storage options for over 25 NIH funded research groups.

Four Aims of the NIH G20 in brief

COMPLETED Aim 1: Upgrade the power supply five-fold

- Upgraded electrical from 208V 300kVA to 4160V 1500kVA and power to 900kW (450kW redundant).
- Installed cooling infrastructure with potential capacity of 300 tons.

COMPLETED Aim 2: Implement a facility-wide modular uninterruptible power supply (UPS)

- Implemented facility-wide UPSs.
- Distributed to power distribution units (PDUs) in the server room.
- Interconnected to overhead power distribution busway (NYSERDA funded).
- Built the infrastructure to connect to backup generator

COMPLETED Aim 3: Double the capacity of the current pilot shared HPC research cluster

- The pilot shared HPC research cluster (Hotfoot) had **616** cores. The SRCF currently has three shared HPC clusters totalling **18,716** cores.

COMPLETED Aim 4: Establish a pilot professionally administered data storage, also funded by a New York State matching grant.

- Launched research storage pilot consisting of 100 TB NetApp
- Created the Secure Data Enclave (SDE) in 2013

Planning for the Future

- On Prem vs. Supercomputing Center vs. Cloud
  - Data Center Capacity
  - TACC
  - Cloud

- Institutional support and funding models
  - F&A on Cloud

Data Center Capacity

- Capacity is measured by space, cooling, and power
- The G20 expanded all three
- In 2018 we hit a cooling constraint, which was solved by expanding the cooling capacity and adding 16 HD racks and 10 cooling units
- Theoretical electrical capacity in the data center is 400kW for HPC
- Current HPC installed footprint is 13 racks, ~200kW
- We retire old equipment as it reaches end of life, rotating equipment within the 16 racks
- The next constraint is power

Rack Utilization

Rack	Loc	2022	2023	2024	2025	2026	2027
1	L18	HABA1	BURG3 (estimate)	BURG3 (estimate)	BURG3 (estimate)	BURG3 (estimate)	BURG3 (estimate)
2	L19	HABA1	BURG3 (estimate)	BURG3 (estimate)	BURG3 (estimate)	BURG3 (estimate)	BURG3 (estimate)
3	L21	HABA1					
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	GPU MQ (9/22)	GPU MQ (9/22)	GPU MQ (9/22)	GPU MQ (9/22)	GPU MQ (9/22)	
15	L27	GPU MQ (9/22)	GPU MQ (9/22)	GPU MQ (9/22)	GPU MQ (9/22)	GPU MQ (9/22)	
16	L28		GPU 2 (estimate)	GPU 2 (estimate)	GPU 2 (estimate)	GPU 2 (estimate)	GPU 2 (estimate)

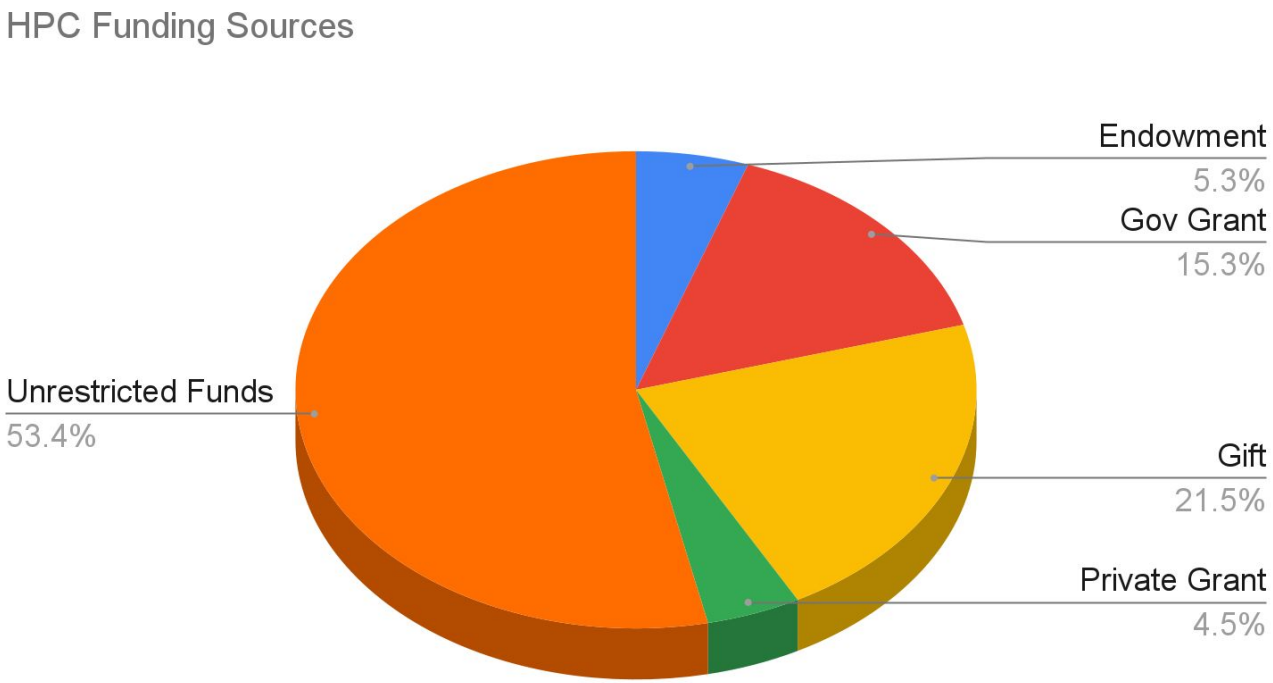
Texas Advanced Computing Center (TACC)

- Designs and operates some of the world's most powerful computing resources.

- 2019 quote for Columbia partnership
  - $1M - for one year
    - ~70M core hours
    - ~125TB storage
    - ~4PB archives storage
    - ~1000 hours programmer/consultant time

(on prem HPC, all clusters for the year of 2021 = 79,279,196 core hours, approximate hardware spend = $1.06M)

Where is the funding coming from?



The Future of Research Computing

CASC - Coalition for Advanced Scientific Computation - 2021 Report

- CASC plans to continue this survey on an annual basis
- 69 Institutions responded

“On-premises delivery of RCD resources remains by far the preferred method in terms of ROI at the vast majority of responding institutions.”

“This year, however, for the first time, a few institutions indicated that commercial cloud resources had become a useful form of delivery for at least some situations their research community and a small number of respondents indicated cloud resources to be better in terms of ROI than fully on-premises delivery in at least some circumstances.”

Reference: Sharon Broude Geva, Alan Chalker, Curtis W. Hillegas, Donald Petravick, Alan Sill, and Craig A. Stewart. 2021. Results from a second longitudinal survey of academic research computing and data center usage: expenditures, utilization patterns, and approaches to return on investment. In Practice and Experience in Advanced Research Computing (PEARC ’21), July 18–22, 2021, Boston, MA, USA. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465589

Other Institutions On-prem

Institution	Funding Model	Cores
Columbia	Condo	18,716
Stanford	Condo	41,250
University of Arizona	unknown	41,444
Princeton	Central funding?	>82,500
NYU Greene	unknown	>59,000
Rutgers	Condo	21,000

Reference: Sharon Broude Geva, Alan Chalker, Curtis W. Hillegas, Donald Petravick, Alan Sill, and Craig A. Stewart. 2021. Results from a second longitudinal survey of academic research computing and data center usage: expenditures, utilization patterns, and approaches to return on investment. In Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465589

F&A Waived for Cloud Computing resources

Educause Review 2015: **Federal Indirect Costs Affect Total Cost of Ownership**

Some institutions waive F&A on Cloud Computing

- University of Illinois
- University of Washington
- UC San Diego
- Georgia Tech

Chair of SRCPAC

Chris Marianetti has been Chair of the SRCPAC Committee since 2016!

We thank him for his extended service to the Shared Research Computing program. He has been an engaged and enthusiastic chair, leading the direction for the successful program, and we will miss him.

- *CUIT Research Computing Services*

Questions?

Supplemental Slides

The Future of Research Computing

CASC - Coalition for Advanced Scientific Computation - 2021 Report

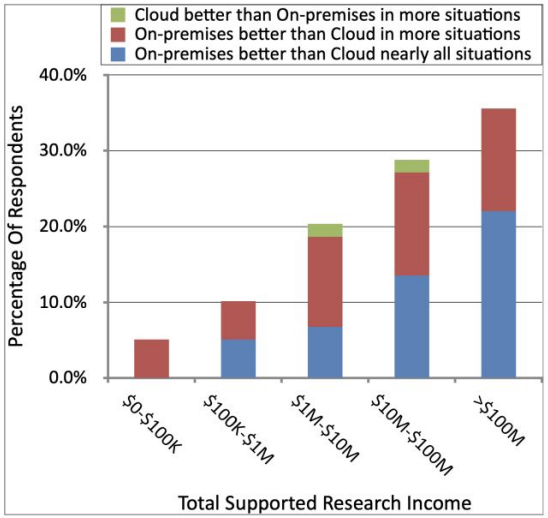


Figure 2: Percentage of total respondents to the 2021 survey to the question regarding assessment of ROI depending on the mode of delivery and level of supported research income.

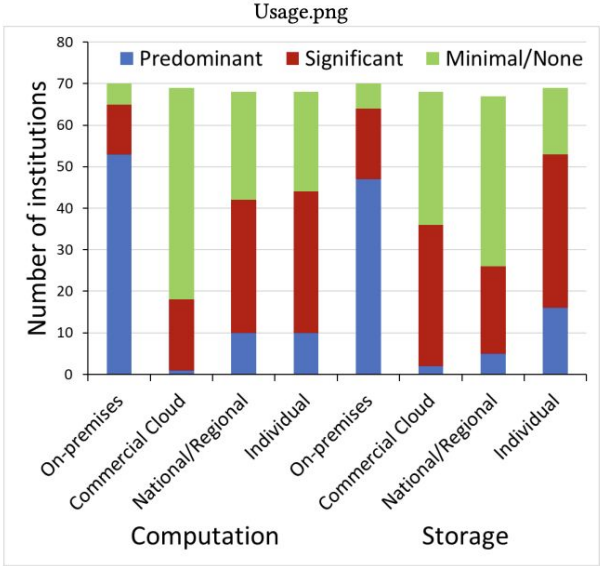


Figure 1: Relative level of usage by researchers organized by mode of delivery.

Reference: Sharon Broude Geva, Alan Chalker, Curtis W. Hillegas, Donald Petravick, Alan Sill, and Craig A. Stewart. 2021. Results from a second longitudinal survey of academic research computing and data center usage: expenditures, utilization patterns, and approaches to return on investment. In Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465589

The Future of Research Computing

CASC - Coalition for Advanced Scientific Computation - 2021 Report

Table 2: Funding models for Computing Resources

What funding models are used for capital resources/operational resources (staff and support) for RCD? (Select all that apply)	Capital	Operational
The institution makes one-time or intermittent investments in on-premises resources	36	n/a
The institution provides a regular budget for on-premises resources	35	37
Colleges or departments fund on-premises resources that are offered to researchers within their own organizational unit	39	29
Researchers or research groups write grant requests to fund resources	55	25

Reference: Sharon Broude Geva, Alan Chalker, Curtis W. Hillegas, Donald Petravick, Alan Sill, and Craig A. Stewart. 2021. Results from a second longitudinal survey of academic research computing and data center usage: expenditures, utilization patterns, and approaches to return on investment. In Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465589

Research Computing Executive Committee

May, 31 2022

Marc Spiegelman, Chair of Advisory Committee
Arthur D. Storke Memorial Professor Department of Earth & Environmental Sciences
Chair: Department of Applied Mathematics and Applied Physics

Outline

- Brief overview of Foundations: Goals, Design, Organization
- Key Components of Foundations
- Accomplishments
- Challenges
- Discussion

What is Foundations for Research Computing?

- University-wide program providing **informal training** in fundamental computational skills for research computing.
- Targets **graduate students and postdocs**
- Initial Priority to **Arts and Sciences**, **Engineering**
- Helps foster **community** for research computing

Design of Foundations

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

Foundations Program Components

Training Opportunities

- **Bootcamps** - 2-4 day training based on *Software Carpentry* curriculum for novice learners
- **Intensives** - 1 day training for intermediate learners with curriculum developed internally or with external partners, e.g. Google
- **Workshops** - 1.5 - 2 hour training to advance computational skills in a group setting. Range of ~10-100 attendees per workshop.

Community Building

- **Python User Group** - grass-roots community computational group for all students and postdocs.

Support/Organization

- Originally envisioned as a 3-year pilot project starting Fall 2018
- First two-years were funded by generous contributions from the Libraries (personnel) and EVPR, A&S and SEAS (operating)
- Since 2020: the Libraries, very generously has assumed both personnel and operational costs.
- Key personnel: the Foundations Coordinator position
  - Was Patrick Smyth (2019-2021)
  - Currently seeking new Program Coordinator/Manager

Foundations for Research Computing: Bootcamps

Bootcamp Content:

- Unix Shell
- Version Control with Git
- Introduction to Python
- Introduction to R

Hands on Live-coding pedagogy: (modified for online due to covid)

- 30 students Maximum per class
- 1-2 Trained Carpentries Instructors - Using well vetted, open source lessons
- 2-4 “Helpers”

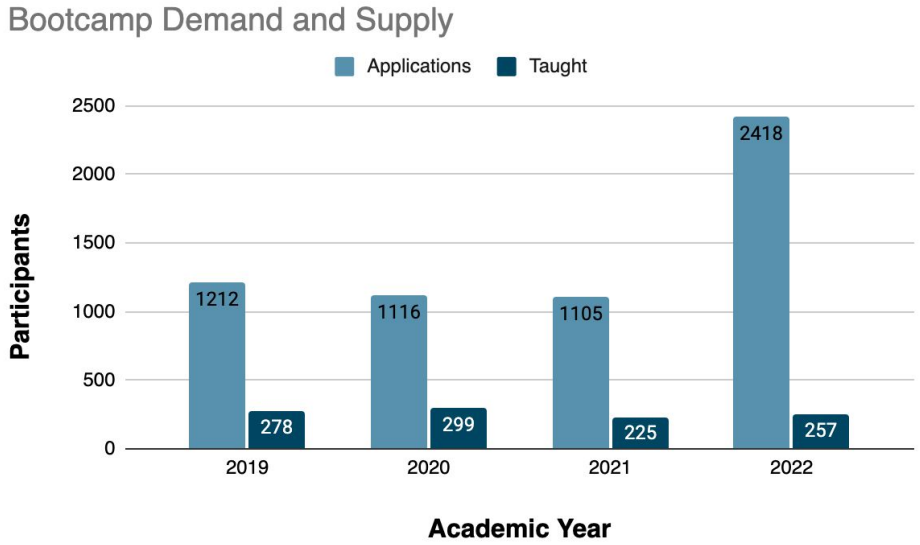
Instructors:

- Since 2018 CU has provided 47 SC trained instructors
- 30 currently active
  - ⅔ Permanent staff (CUIT, Libraries, sys-admins etc)
  - ⅓ Graduate students
  - all volunteers

Foundations for Research Computing: Bootcamps

software carpentry

Teaching basic lab skills for research computing



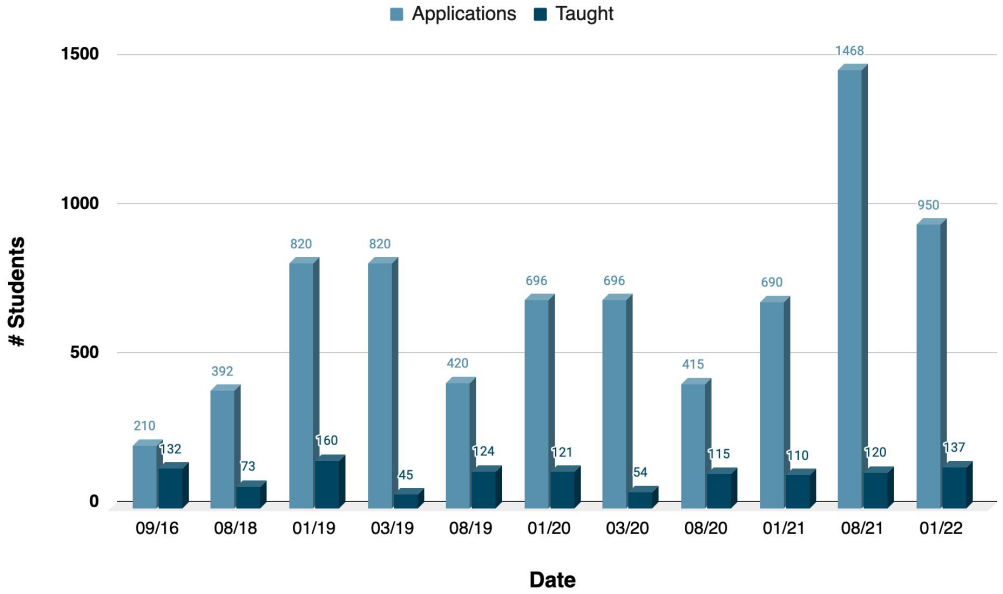
Current Demand for positions is ~10x available

Foundations for Research Computing: Bootcamps

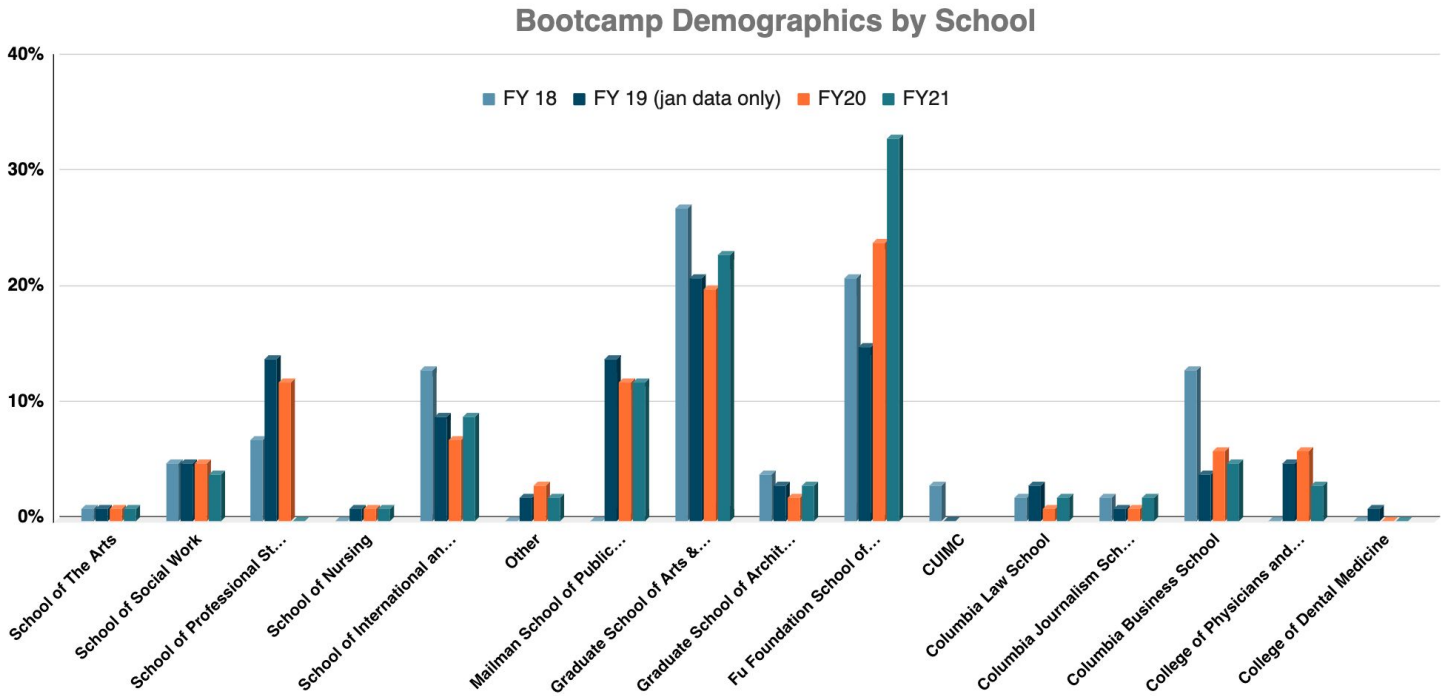
software carpentry

Teaching basic lab skills for research computing

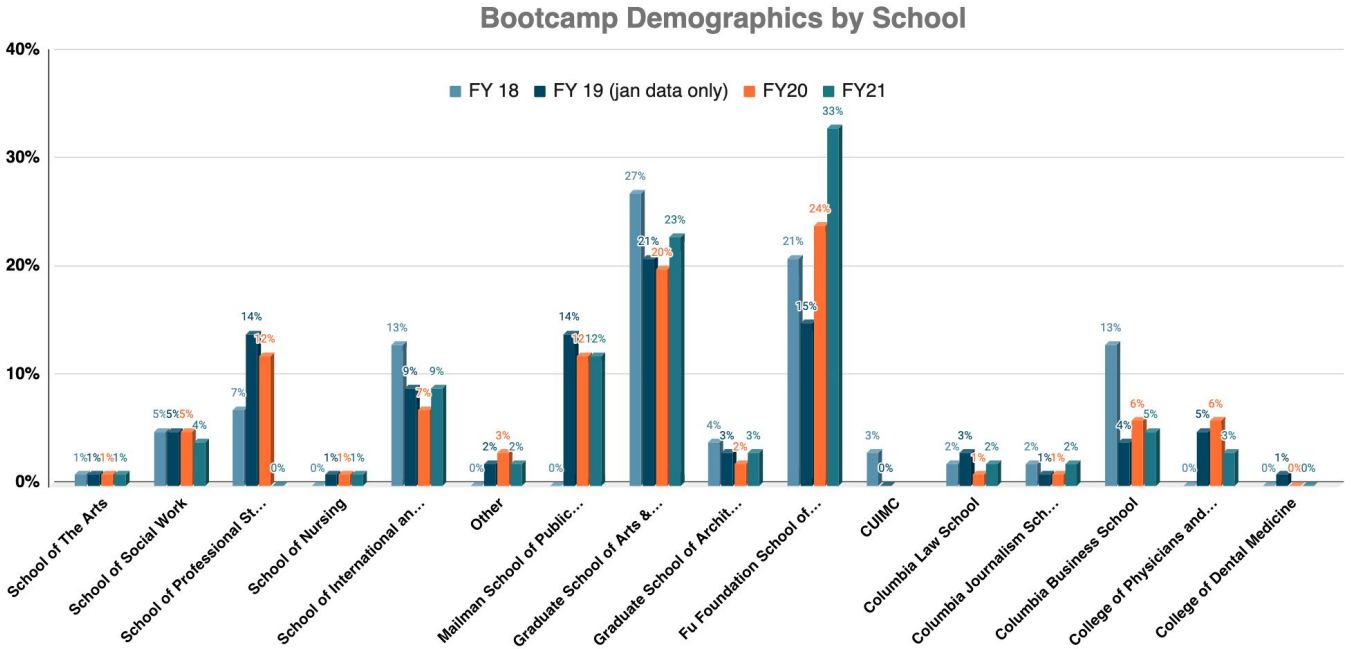
Current Demand is ~10x available positions



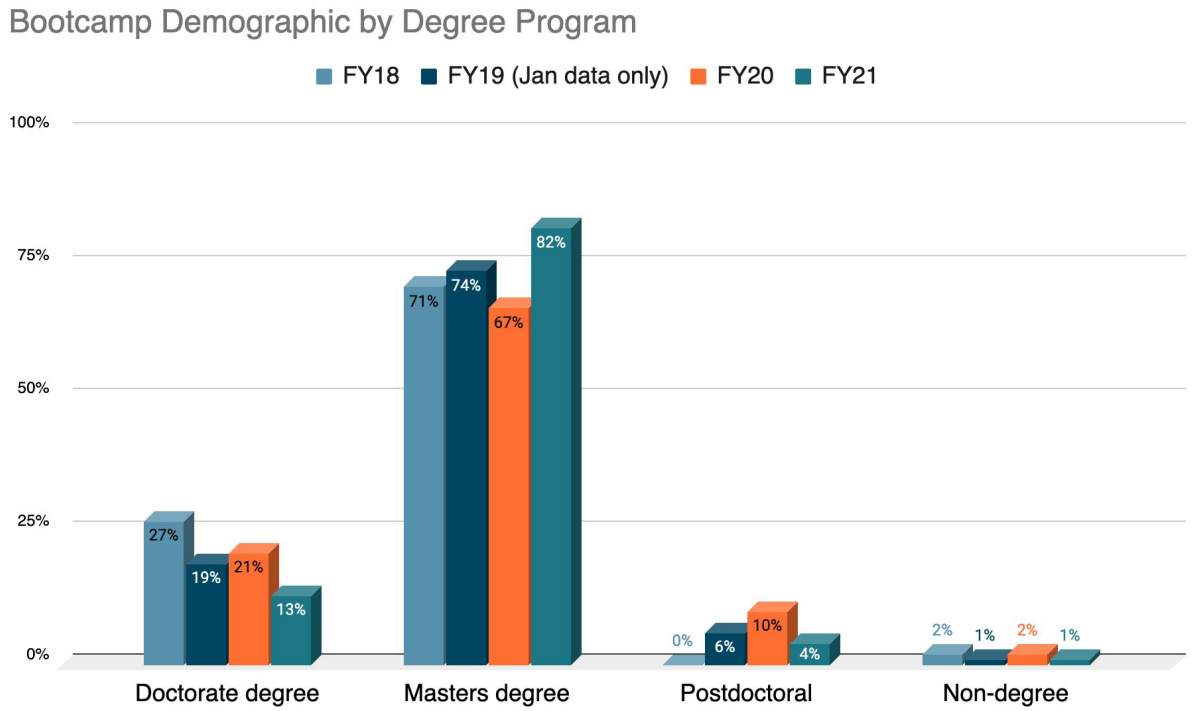
Foundations for Research Computing: Bootcamps



Foundations for Research Computing: Bootcamps



Foundations for Research Computing: Bootcamps



Foundations for Research Computing: Intermediate

- **RCS Workshops** (AY2021 - 10 workshops - 379 participants)
  - Introduction to Linux
  - Introduction to Scripting
  - Introduction to HPC
  - Introduction to Cloud Computing
- **Foundations and PUG Workshops** (18 Workshops - 447 participants)
  - Tensorflow 2.0 (w Google)
  - Pandas
  - PyMC3 for probabilistic programming
  - Text Analysis with SpaCy

Python User Group - 2021/2022

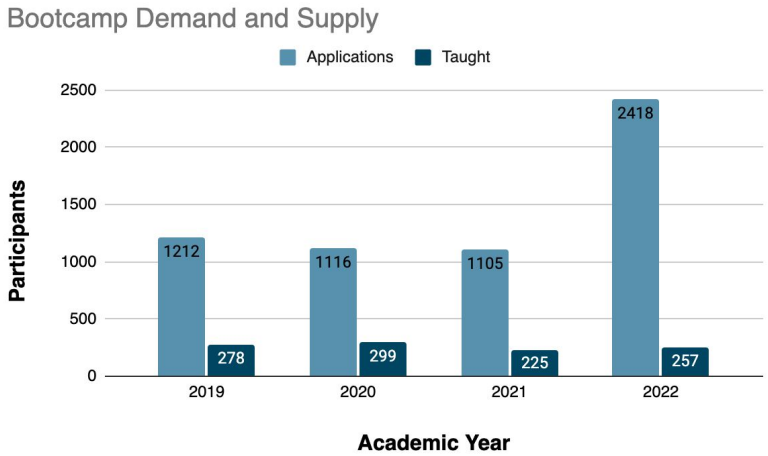
Python User Group - meets twice-monthly, hired two graduate students who work closely with a member of the Libraries Research Data Services to identify topics and develop curriculum.

- Scaling Python Analytics with Dask
- General Intro to Pandas & EDA
- Intro to Julia for Python Users
- Geopandas & You
- Intro to Text Analysis in Python
- Test-Driven Development & ML Audio Transcription

Foundations for Research Computing: Demand

Foundations is engaged in a wide variety of activities at Novice and Intermediate level reaching ~1000 participants/year

But demand has always been higher than what Foundations can provide.



Foundations for Research Computing: Challenges

Continuing questions:

- How to scale to meet demand?
  - Horizontally: Increasing participation in Novice bootcamps
  - Vertically: How to increase offerings for more advanced students?
- How to articulate different training needs among these applicants?
- Who should Foundations serve?

Models for Expansion: Departmental Partnerships

Mechanical Engineering

- Ran a Software Carpentry bootcamp in mid August as part of incoming masters student orientation
- Adapted Python portion to be discipline specific
- Intends to offer a full week bootcamp next year
- Successfully running smoothly, minimal central resources
- But leverages Foundations trained instructors

Models for Expansion: Departmental Partnerships

Social Sciences Bootcamp: (AY 2020)

- Trained Software Carpentry instructors from Psychology Department
- 46 grad students and postdocs, 39 others
- Advertised on Foundations for Research Computing listserv & website
- Three day-long intensives:
  - Research Computing for Social Scientists
  - R for Social Sciences Data
  - Python for Social Sciences Data
- Discipline-specific curriculum developed & iterated on by psychology department

Models for Expansion: Supporting Center Grant Proposals

LEAP STC: Learning the Earth with Artificial Intelligence & Physics

- Participated in successful proposal for a $25 million NSF Science & Technology Center award
- Currently working with Tian Zheng, *Chair, Department of Statistics & Education Director for LEAP* to train students as Software Carpentry instructors, 1-2 per semester
- Depending on timing, they will participate as helpers for the regular bootcamps, before being instructors for the LEAP bootcamps
- Those students will join the growing Foundations instructor community on campus

Summary

- Foundations is providing much needed training support for research computing at Columbia
- Financially sustainable with generous Library support
- However, demand is insatiable…
- Experimenting with new methods for increasing participation and tailoring to specific CU needs

Possible Future Directions

- Engage with Departments to see viability of expanding/regularizing MechE model
- Expand & Nurture SC Instructor pool
- Develop more on-demand asynchronous offerings
- Develop mechanisms to provide better integration with Departmental Course offerings.

All of these require a dedicated Foundations Program Manager (search in progress)

But now is a good time to reassess needs and directions. An extended advisory board will be meeting shortly to discuss faculty perspectives re: Foundations.

Discussion

End of presentation

Year 2 Overview: Curriculum Innovation Grants

An Experiment

- Awarded to **graduate students** and **postdocs** to create a learning module on a particular aspect of research computing of use to the Columbia Community
- Includes
  - Training as a Software Carpentries Instructor
  - Collaboration with Center for Teaching and Learning
  - A small stipend
- 5 funded through Foundations + 2 through QMSS

Year 2 Overview: Curriculum Innovation Grants

Status

Topic	Workshop?	Event link	Curriculum?	Curriculum link
Data Manipulation and Visualization in R	Yes	Link	No	
Intro to Deep Learning with PyTorch	No		No	
Python for the Analysis and Visualization of Biological Datasets	No		No	
Tidying Survey Data in R	Yes	Link	No	
Data Analysis and Manipulation with Xarray	Pending		Pending	
Interactive Data Visualization in R and Shiny	Yes	Link	Yes	Link
Wrangling Multilevel Data in the R Tidyverse	Yes	Link	Yes	Link

Year 3 Planning: Overview

- Training: ~600 students and postdocs in a mix of
  - Standard Bootcamps, Intensives, workshops (possibly hybrid or remote with guidance from Carpentries)
  - Expanded disciplinary pilots (CUIMC, MechE, Humanities)
- Curricular Innovation Grants (3)
- Python User Group
- Distinguished Lectures on hiatus

Year 3 Proposed Disciplinary Pilots

Three Disciplinary Pilots

Build capacity within a discipline by training Software Carpentry instructors from

- Mechanical Engineering
- CUIMC
- Humanities

Year 3 Disciplinary Pilot: CUIMC

Three potential opportunities to expand our reach at CUIMC

- Train **3rd year medical students** (30-40) working in the Scholarly Projects Program.

  Contact: Bill Bulman, Associate Professor of Medicine, CUIMC

- Train **Fellows** (40-50), post residency students sub-specializing and working closely with research intensive faculty.

  Irving Institute for Clinical and Translational Research
  Contact: Muredach Reilly, Director Irving Institute

- Plan for a train the trainer model, we provide infrastructural and light administrative support

  Contact: Art Palmer, Robert Wood Johnson, Jr. Professor of Biochemistry and Molecular Biophysics

Year 3 Disciplinary Pilot: Humanities

Opportunity: Experiment with different models for domain specific training, expanding the reach of the program beyond STEM

Approach

- Create a modular based, humanities intensive for graduate students and postdocs in the humanities
- Explore training graduate students in SC Pedagogy for teaching opportunities in year 4

Contacts: Dennis Tenen, Associate Professor of English & Comparative Literature and Manan Ahmed, Associate Professor of History

Summary

- Program is expanding to meet demand (both numbers and content)
- Experimenting with new methods for increasing participation and tailoring to specific CU needs
- Financially sustainable with generous Library support
- We just need to remain agile (like everyone else) and respond to the times.
- But these are some of the skills that are still useful in a socially distanced universe

Discussion

Guidance on partnerships w/schools & depts to expand/extend beyond PhDs and into discipline specific training?

- Pilot was to prioritize PhDs & Postdocs, includes Master students when there is room
- Demand to include masters and discipline specific
- Proposed solution is to partner with schools & departments to train the trainers (PhD & postdoc)
- Piloting with 3-5 schools/departments this upcoming year
- Working on right financial model