COLUMBIA UNIVERSITY

Update to SRCPAC, Spring 2021

Dear SRCPAC Members,

SRCPAC traditionally meets every Fall and Spring. As we again have no pressing issues, in lieu of yet another Zoom meeting, attached please find updates on:

- Our High Performance Clusters as well as some information on cloud computing resources and electronic lab notebooks. The graph showing the core growth from our initial HPC Hotfoot really impresses! Ginsburg – welcome to the family. Habanero – glad to see you made the move largely intact.
- Foundations for Research Computing. Even under COVID constraints, between bootcamps, workshops and the Python User Group, Foundations has reached over 1000 researchers this year.

If all goes well, we shall look forward to meeting in person this Fall. In the meantime, should you have any questions or comments, please feel free to email srcpac@columbia.edu.

High Performance Computing

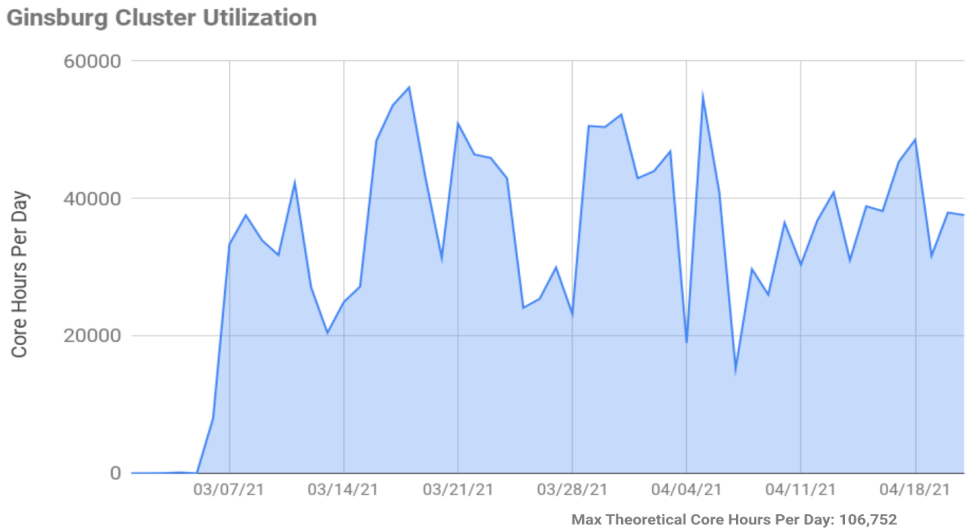
GINSBURG

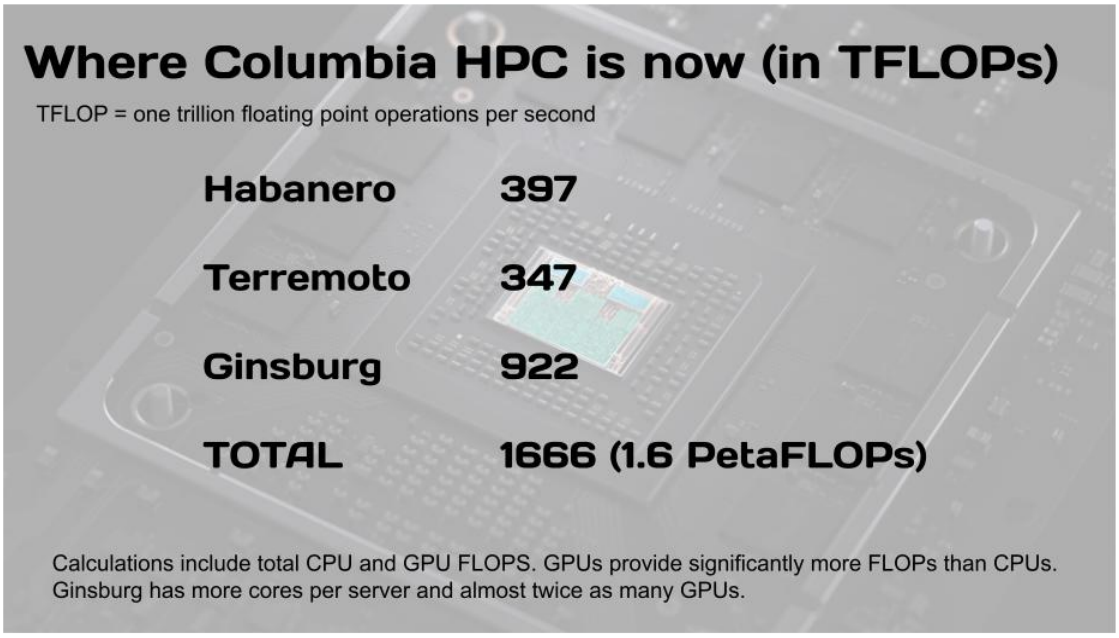
The Ginsburg cluster went live and became available to researchers in February 2021. This newest cluster is now 139 high-throughput compute nodes:

- ❑ 87 standard nodes
- ❑ 30 high memory nodes
- ❑ 22 GPU nodes
- ❑ 570TB storage

We will begin our expansion purchase round starting in Spring 2021 for November 2021 go-live.

Ginsburg researchers have already utilized over 1.7 Million core hours since the system went live less than 2 months ago.





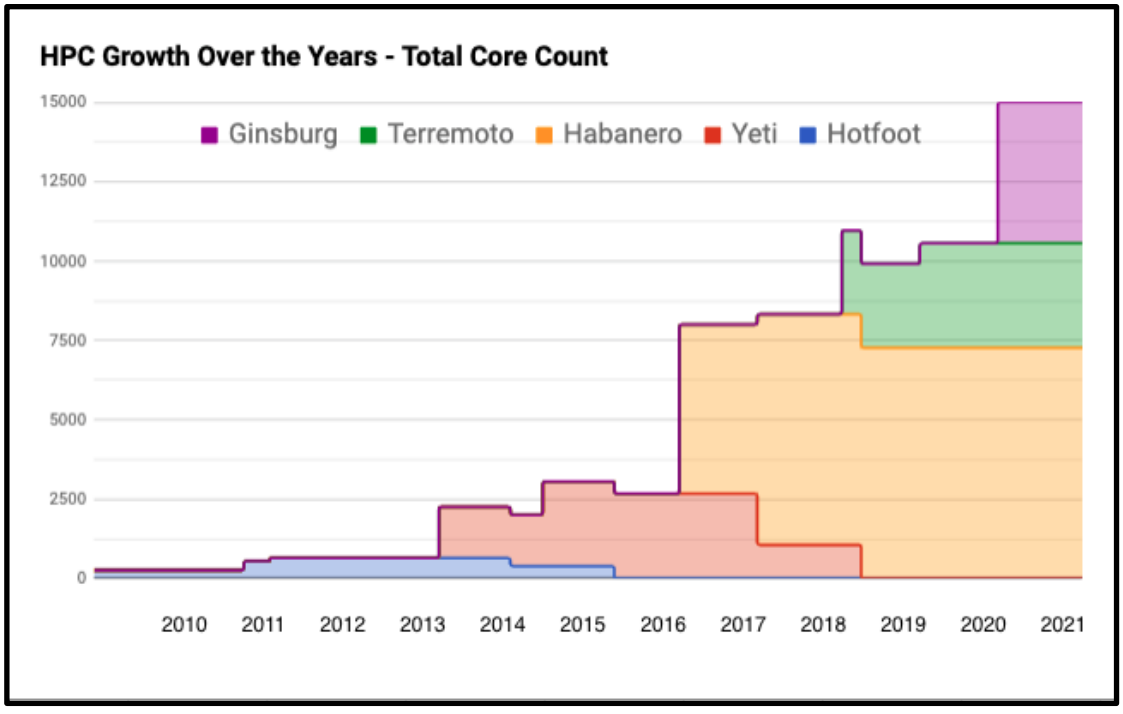
HABANERO

In March 2021, the CUIT Research Computing Services team successfully relocated the Habanero cluster from the Zuckerman Data Center to the CUIT Computer Center.

This was quite an effort, but the RCS team had Habanero back up and running in less than two weeks' time.

If you elected to have your nodes moved, you can now enjoy another year on our most popular cluster. You will see the $250 charge per node as 'services provided by CUIT' on the chartstring you provided. You will continue to have the same rights and priorities. We will start the year extension, effective the date of the move and we plan to retire Habanero April 1, 2022.

Fun fact: All three current high performance computing clusters, **Ginsburg**, **Habanero**, and **Terremoto** are now housed in the **CUIT Computer Center**. This was made possible by the computer center improvements made over the past years with contributions from A&S, SEAS, EVPR and CUIT.



More CUIT RESEARCH SERVICES

Google Cloud Platform (GCP) for Research

During the Fall and Spring semesters, RCS developed and offered two training courses, geared towards helping researchers understand how to leverage GCP resources for their research:

- ❏ Introduction to Cloud Computing for Research
- ❏ High Performance Computing on GCP

These virtual classes were well-attended and we will continue to offer them regularly. See the Research Computing Services (RCS) for scheduling information.

Columbia now has BAAs with Google Cloud Platform (GCP), Amazon Web Services (AWS), and Microsoft Azure (Azure). CUIT recommends GCP as the preferred platform for research computing, as we are best positioned to provision and provide support for GCP.

RCS Consulting

As a reminder, you can consult with CUIT's **Research Computing Services** team (**RCS**) on your research computing needs. Whether you have questions about cloud computing, on-prem HPC, or external resources such as XSEDE, please reach out to rcs@columbia.edu.

Electronic Research Notebooks with LabArchives

Columbia University provides an Electronic Research Notebook service for researchers, instructors, and students, designed to replace paper notebooks and lab manuals to support research staff productivity and efficiency, and securely protect research with automatic backups and comprehensive audit trails. Our Electronic Research Notebook solution is powered by LabArchives, a cloud-based solution that enables 24/7 collaboration and sharing of content, from any device with a web browser.

Note that LabArchives has now been certified by CUIMC Security IT for PHI.

Electronic research notebooks also **support funding agencies' Data Management Plan** requirements, manage team progress remotely, and interconnect all your data and image files to your observations and notes.

Columbia University has an enterprise license with LabArchives for both the Professional Edition and Classroom Edition of their electronic notebook solution.

Columbia's Electronic Notebook Service is funded by CUIT and the Libraries, in partnership with the Office of the Executive Vice President for Research, and is provided at **no cost to you.**

Foundations for Research Computing

In spring, 2021, the Foundations for Research Computing program continues to offer a diverse array of training opportunities for researchers at Columbia to develop and improve computational skills. Foundations is a collaboration among CUIT, the Office of Research, and Columbia Libraries, and offers a variety of events and training to build capacity and community around computational research at Columbia. In the past months, the program has experimented with new event formats to engage researchers remotely, including trials of a four-day Foundations for Research Computing bootcamp and a set of more flexible mid-semester offerings at the intro level. The program has continued its pilot partnerships with university departments, providing targeted resources and guidance for a two-day bootcamp for graduate students in the department of Cardiology in January. The program has also seen an increase in interest for mid-semester training, and Foundations for Research Computing is on track to engage significantly more researchers than projected in the summer, 2020 presentation to RCEC.

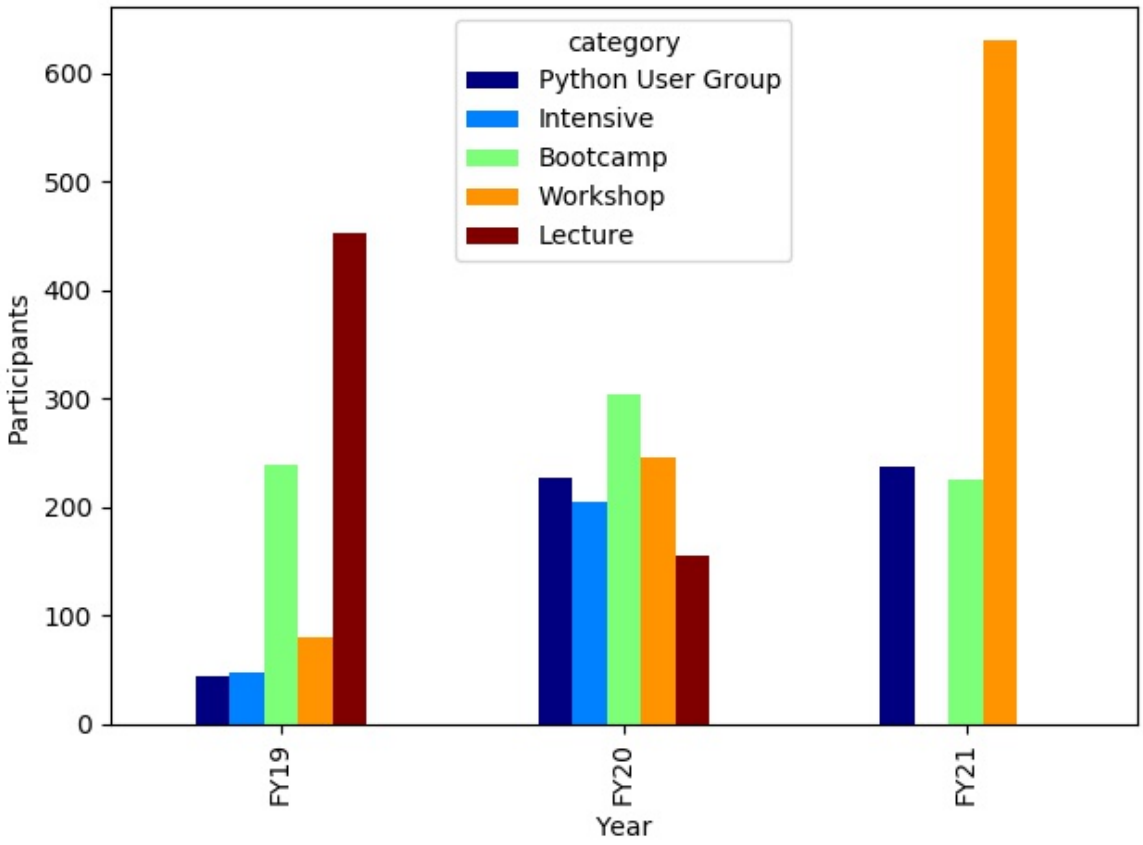
As this is a longer update, we have broken it into three sections to focus on specific areas of interest:

- Spring Programming
- Event Formats
- Partnerships

Programming

Since January, 2021, Foundations for Research Computing has seen an uptick in researcher interest in Foundations programs other than the bootcamps. For the first time, subscribers to the primary Foundations for Research Computing mailing list have exceeded 4000. Mid-semester events now frequently begin to fill up before they are actively promoted, since many researchers now check the program calendar on a regular basis. This semester, Foundations has organized a diverse set of workshops at all levels, from two workshops on Getting Started with Python to an introduction to functional programming using the Haskell language. In March, Foundations was able to organize a larger workshop on machine learning framework TensorFlow 2.0 by Josh Gordon of Google AI.

CUIT's Research Computing Services (RCS) has been offering internally-developed workshops on Linux, scripting, and High Performance Computing and extends registration to the Foundations audience. This year, RCS added two new workshops, Introduction to Cloud Computing for Research and High Performance Computing on Google Cloud Platform. This research computing series, now consisting of five workshops, was notably popular, with over 81 participants at the Introduction to Cloud Computing for Research workshop and 51 at High Performance Computing on GCP.



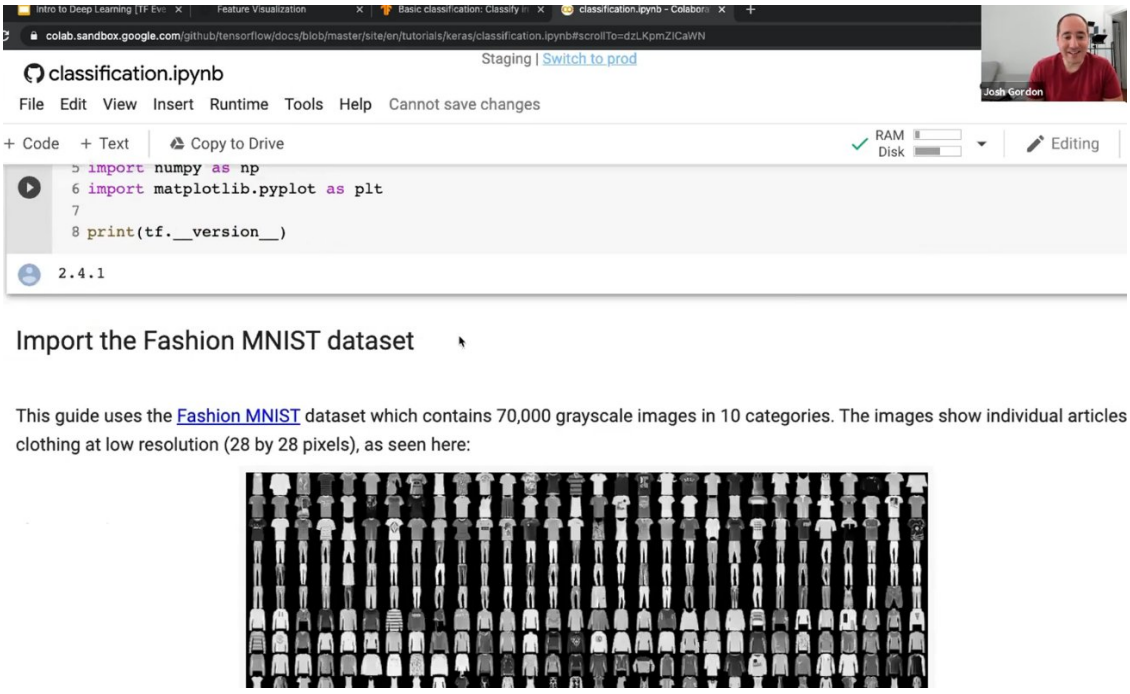
Despite necessary reduction in some categories, the program has expanded workshop offerings, and categories such as bootcamps and Python User Group remain stable through challenges. Workshops in this chart include the CUIT offerings.

Since February, Foundations has engaged 583 participants between workshops and Python User Group over 13 events. In the year to date, and including bootcamps, workshops, and Python User Group, Foundations reached over 1054 researchers, exceeding the 500 estimate made to RCEC in summer, 2020. Last year, the program served 1136 researchers, so despite substantial reductions in some areas of the program, expansion in areas such as workshops mean that Foundations is on track to serve approximately as many researchers as in Fy 2020.

Event Formats

In January, 2021, following new guidance from The Carpentries on lowering Zoom fatigue during remote training, the program experimented with spreading the usual two-day bootcamp over four sequential mornings. The event was also divided into two tracks to reduce the overall number of participants in the Zoom room to reduce distraction and facilitate greater individual attention. Feedback on the new format from participants was mixed. Though it is difficult to make direct comparisons to prior events, the four-day format may have increased, rather than decreased, fatigue in participants, and a number of participants made suggestions for future events related to changes in format.

Based on feedback from participants, and continuing in a spirit of improvement through iteration, the planned spring break bootcamp was replaced with a spaced-out and flexible series of introductory workshops in March and early April. Prior iterations of a March bootcamp have seen something of a "spring break effect," and we have learned that this week is a less desirable time for a concerted focus on new skills than January or August. Preliminary feedback has suggested this may be a viable augmentation or supplement to our usual introductory programming. A determination has not yet been made as to the format for the program's August bootcamp, and the program may modify the four-day format or return to a two-day format.



Josh Gordon of Google AI leads a session on machine learning framework TensorFlow

---

Partnerships

In these ongoing pilot partnerships, Foundations for Research Computing provides initial consultation, a packet of checklists and communications, and a sponsorship for training by The Carpentries. Partners use these resources to run an event or events for their own department, center, or school. These partnerships are designed to build technical and instructional expertise in computational research around the university by leveraging existing program infrastructure. Partners are often more acutely aware of local needs for technical expertise, and the program hopes that these partnerships will facilitate the teaching of more customized events for specific disciplines and groups.

Following a successful August, 2020 cooperation with Mechanical Engineering, the program continued to pilot new partner models. In preparation for a course on statistics taught by Art Palmer in the Division of Cardiology, the program assisted Roger Lefort of Research Compliance and Training in the Office of Research as he led a Software Carpentry bootcamp for 25 PhD students.

One insight from conducting this pilot has been that seeding new events and training with one or two of our experienced instructors can greatly assist a department or group with running an event for the first time, as working with experienced instructors is a learning opportunity for those new to event organization or technical training. The program is currently exploring the possibility of formalizing this work by designating "senior instructors" who are available to work

with departmental, school and institute partners in this capacity. The program is also considering ways to formalize partnership models, which would appear on the Foundations for Research Computing website and which would serve as a starting place for discussions with new departments, laboratories, and groups.

In spring, 2021, Foundations continues to explore new ways to engage Columbia researchers to build a community around computational research at the university. Thank you to SRCPAC for the committee's continued support. Please reach out at any time with questions, either to rcfoundations@columbia.edu or to the Foundations for Research Computing program coordinator at patrick.smyth@columbia.edu.

Events Held in FY2021

Date	Name	Engagements	Initiative Category
8/24/20	Foundations for Research Computing Bootcamp	115	Bootcamp
1/11/21	Foundations for Research Computing Bootcamp	110	Bootcamp
9/8/20	Python Fundamentals for Data Cleaning	29	Community
9/22/20	Python User Group: PyMC3 for Probabilistic Programming	37	Community
11/3/20	Python User Group: Make Your Code Interoperable by Making a CLI Program	9	Community
10/20/20	Python User Group: Data analysis and manipulation with Pandas	15	Community
10/26/20	Reading Group: GPT3	5	Reading Group
9/28/20	Text Analysis in Python with SpaCy	29	Workshop
9/29/20	(CUIT) Introduction to Linux	20	Workshop
10/6/20	(CUIT) Introduction to Scripting	21	Workshop
10/13/20	(CUIT) Introduction to High Performance Computing	20	Workshop
10/19/20	Introduction to Functional Programming with Haskell	14	Workshop
10/20/20	Python User Group: Data analysis and manipulation with Pandas	16	Workshop
10/26/20	Python for Beginners	13	Workshop
10/28/20	Python for Beginners	13	Workshop
11/12/20	Data Cleaning In the Face Of Data Issues: Exploring the NYPD Misconduct Complaint Database	13	Workshop
11/17/20	(CUIT) Introduction to Cloud Computing for Research	19	Workshop
11/19/20	(CUIT) High Performance Computing on Google Cloud	17	Workshop

	Platform		
2/9/21	Pandas: The Bare Basics	71	Community
2/22/21	Text Analysis with Haskell	47	Workshop
2/23/21	Python User Group: Extracting Data from APIs	46	Community
3/8/21	Dive into TensorFlow 2.0	45	Workshop
3/9/21	Python User Group: Working with the Google Sheets API in Python	31	Community
3/15/21	Make a Webpage	32	Workshop
3/16/21	(CUIT) Introduction to Cloud Computing for Research	81	Workshop
3/16/21	Getting Started with Python	20	Workshop
3/18/21	(CUIT) High Performance Computing on Google Cloud Platform	51	Workshop
3/23/21	(CUIT) Introduction to Linux	53	Workshop
3/24/21	Getting Started with Python	9	Workshop
3/30/21	(CUIT) Introduction to Scripting	53	Workshop
4/6/21	(CUIT) Introduction to High Performance Computing	44	Workshop