Shared Research Computing Policy Advisory Committee (SRCPAC)
Spring 2020 Meeting
April 16, 2020

Meeting Called By: Chris Marianetti, Chair of SRCPAC
SRCPAC Agenda

● Welcome & Introductions
  ○ Chris Marianetti, Chair of SRCPAC

● High-Performance Computing Update
  ○ Kyle Mandli, Chair of the HPC Operating Committee
  ○ George Garrett, Manager of Research Computing, CUIT

● Foundations for Research Computing Update
  ○ Marc Spiegelman, Chair of the FoRC Advisory Committee
  ○ Barbara Rockenbach, Associate University Librarian for Research and Learning

● Other Business & Closing Remarks
  ○ Chris Marianetti, Chair of SRCPAC
High Performance Computing Update

Kyle Mandli, Chair of the HPC Operating Committee
George Garrett, Manager of Research Computing, CUIT
HPC Agenda

- Governance
- COVID-19 HPC Update and Consortium
- HPC Cluster Stats and Updates
- HPC Expansion Update
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 Fall 2020.
- All HPC Users (Terremoto, Habanero) are invited to participate.
Columbia researchers are conducting COVID-19 research in many different areas. A couple examples of researchers using Columbia HPC machines for COVID-19 research are:

- **Stockwell Lab**: Working on SARS-CoV-2 protease inhibitors with David Ho and Alex Chavez
- **Przeworski Lab**: Modeling and analysis of phylogenetics of available genomes

**COVID-19 High Performance Computing Consortium**

- [https://covid19-hpc-consortium.org/](https://covid19-hpc-consortium.org/)
- Bringing together the Federal government, industry, and academic leaders to provide *free access to the world’s most powerful high-performance computing resources* in support of COVID-19 research
- Researchers are invited to submit COVID-19 related research proposals to the consortium (418 Petaflops of total processing power available)
Habanero and Terremoto - Participation and Usage

- 44 groups
- **2050** users since launch (265 active)
- 22 renters since launch
- 395 free tier users since launch
- Education tier
  - 22 courses since launch (4 added in Spring 2020)

- 32 research groups
- **512** users (161 active)
  - (187 users added since October 2019)
- **24 million core hours** utilized since launch in December 2018
- Launched in December 2018
- Expanded in December 2019
- 5 year lifetime

- 137 Compute Nodes (3,288 cores)
  - 111 Standard nodes (192 GB)
  - 14 High Memory nodes (768 GB)
  - 12 GPU nodes with NVIDIA V100 GPUs

- 510 TB storage (DDN GPFS GS7K)
Terremoto - Cluster Usage in Core Hours

Max Core Hours Per Day Pre-Expansion = 63,360
Max Core Hours Per Day Post-Expansion = 78,912
Total core hours used since launch: 24 million

Aug - October 2019  70%
Nov - Dec 2019     76%
Jan - Feb 2020     88%
Mar 1 - Mar 15 2020 85%
Mar 16 - Mar 31 2020 90%
Habanero

- Launched in 2016
- Expanded in 2017
- 4 year lifetime

Specs
- 302 nodes (7248 cores)
- 800 TB storage (DDN GS7K GPFS)

Lifespan
- 222 nodes expire December 2020
  - A subset (~ 1/3) of retired nodes will be moved to the main data center and repurposed for Edu and Free tier use in 2021
- 80 nodes expire December 2021
Max Theoretical Core Hours Per Day = 174,528
Total core hours in past 12 months: 41 million
Galileo Evaluation and Cloud Pilot

• Focus on your work, not your computing infrastructure.

• Drag and drop your containerized docker folder to local server or cloud computing resource (AWS, GCP)

• For more info and a 5 minute demo of launching a job with Galileo, see https://galileoapp.io

• **Compute:** Use Galileo to connect to another Galileo-enabled machine or any cloud resource.

• **Run:** Easily set up and deploy your code.

• **Get results:** Track your job’s status and get notified when your results arrive.

Contact us if interested in helping pilot this, or learning more about Galileo.
HPC Purchase Round - Spring 2020

• Announcement of buy-in opportunity will be sent out in late April
• New cluster and new machine types
• Purchase round to commence in mid-May and be open through mid-June 2020
• Go-live of new equipment in late Fall 2020

If you are aware of potential demand, including new faculty recruits who may be interested, please contact us at rcs@columbia.edu
HPC Purchase Round - Menu Options and Specs

All servers come with Cascade Lake Intel 6226R cpus (2 per server)

- 2.9 GHz speed, 32 cores per server
- Servers offer over 50% more total CPU performance than prior generation Terremoto servers

HPC Purchase Options *

- Standard server (192 GB)
- High Memory server (768 GB)
- GPU server (2 x Nvidia V100S)
- GPU server (2 x Nvidia RTX 8000)
- GPU server (2 x Nvidia RTX 6000)

* Final Price estimates to be determined, and are less expensive than Terremoto servers.
HPC Support Services

- **Email**
  - [hpc-support@columbia.edu](mailto:hpc-support@columbia.edu)

- **Office Hours (Online)**
  - Speak with HPC support staff via Zoom from 3pm – 5pm on 1st Monday of month: [RSVP required](#)

- **Group Information Sessions**
  - HPC support staff meet with your group

- **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
SRCPAC

Aprile 16, 2020
Questions for SRCPAC

● How do we prioritize in this moment? If budgetary conditions require a smaller program, what programmatic elements should we retain in the coming year?
● What new audiences should we reach in year 3 of the pilot?
● What emerging tools and methods should we be following, e.g. TensorFlow?
Foundations for Research Computing: Highlights

**Total Students Trained by Year**

- Year 1: 200
- Year 2: 600

**A&S and SEAS Students 2 year total**

- A&S: 200
- SEAS: 150

**Year 2 Goals:**
- 36 Total Events

**Year 3 Goals:**
- 600 - Bootcamps and Intensives
- 900 - All training events

35 Carpentry Instructors Trained Total
Year 2 Overview

- New intermediate intensives (one-day training events) target more experienced researchers
- Curriculum Innovation Grant Fellows create materials and teach workshop in needed areas
- Reached 999 students across the program, increased attendance across categories
- New processes (sign-in, documentation, coffee) and strong feedback in two-day bootcamps
- Further developing on-campus partnerships
Number of Attendees: 2019–20 Academic Year

304 Boot camps
- 124 Fall bootcamp
- 124 January bootcamp
- 56 March virtual bootcamp

204 Intensives
- 78 Accelerated Python
- 63 Tensorflow
- 46 Social Science R & Python
- 17 Tensor Network

336 Workshops & PUG
- 14 workshops
- 12 Python User Groups (PUG)

155 LECTURE ATTENDEES
- 2 Distinguished Lectures
March Bootcamp Online

- Taught 56 students over two days
- The best instructors were so good we needed fewer helpers
- The technology worked surprisingly well
- JupyterHub and Zoom webinar recommended for added functionality
- Getting the best instructors for this format is important
- Use Software Carpentries Tips for Teaching Online
Curriculum Innovation Grant = Curriculum Innovation Grants are awarded to graduate students and postdocs to create a learning module on a particular aspect of research computing

- Interactive Data Visualization with R & Shiny*
- Intro to Deep Learning with PyTorch
- Wrangling Multilevel Data with R & the Tidyverse*
- Data Analysis and Manipulation with Xarray
- Python for the Analysis and Visualization of Biological Datasets
- Tidying Survey Data in R (with support from QMSS)*
- Data Visualization in R (ggplot2) (with support from QMSS)*
Disciplinary Pilot = Building capacity within a discipline by training 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
Disciplinary Pilot = Building capacity within a discipline by training Software Carpentry instructors from Mechanical Engineering Department

- Approach
  - Train MA cohort (120 students)
  - Dedicated bootcamp or sections of bootcamps
  - Train 3-5 SC instructors from Mechanical Engineering
  - Instructors serve as assistants for bootcamps
- Contact: Jeff Kysar, Professor of Mechanical Engineering and Otolaryngology/Head and Neck Surgery, Chair of Mechanical Engineering
Disciplinary Pilot = Building capacity by training Software Carpentry instructors from CUIMC

● Approach
  ● Train 3rd year medical students (30-40)
    Students deepening their research at this point in their education working in the Scholarly Projects Program.
    Contact: Bill Bulman, Associate Professor of Medicine, CUIMC
  ● Training Fellows (40-50)
    Students are post residency and are sub-specializing and working closely with research intensive faculty.
    Irving Institute for Clinical and Translational Research
    Contact: Muredach Reilly, Director Irving Institute
Disciplinary Pilot = Building capacity by training Software Carpentry instructors within the humanities

● Approach
  ● Create a humanities intensive for graduate students and postdocs in the humanities
  ● Explore training graduate students in Software Carpentry for teaching opportunities in year 4
  ● Contacts: Dennis Tenen, Associate Professor of English & Comparative Literature and Manan Ahmed, Associate Professor of History
Program Goals for the 2020-21 Academic Year

Participation Targets:

- 600 students and postdocs trained via bootcamps & intensives
  - Expand disciplinary pilot to CUIMC & MechE
  - Experiment in Humanities
- 900 students and postdocs total trained (including mini-training opportunities)
Pilot Year 3 Program Plan

- 4 bootcamps (3 general; 1 Mechanical Engineering)
- 7 intensives (6 general; 1 humanities)
- 6 Curricular Innovation Grants
- 4 Distinguished Lectures
- Python User Group
- Workshops
If FY21 budgetary constraints require a smaller program, what programmatic elements are critical to retain?

- Bootcamps
- Intensives
- Workshops
- Curricular Innovation Grants
- Distinguished Lectures
- Python User Group
Questions for SRCPAC

● How do we prioritize in this moment? If budgetary conditions require a smaller program, what programmatic elements should we retain in the coming year?
● What new audiences should we reach in year 3 of the pilot?
● What emerging tools and methods should we be following, e.g. TensorFlow?
Questions for SRCPAC

● What new audiences should we reach in year 3 of the pilot? Possibilities include: Mechanical Engineering MAs, CUIMC, and humanities
● What emerging tools and methods should we be following, e.g. TensorFlow?
● If budgetary conditions require a smaller program, what programmatic elements should we retain in the coming year?
End of presentation
Bootcamp = 2 day training based on Software Carpentry curriculum for novice learners

- August 2019 Bootcamp - 124 students
- January 2020 Bootcamp - 124 students
- March 2020 Bootcamp - 56 students - Online!
Events in FY20: Intensives

**Intensive** = 1 day training for intermediate learners with curriculum developed internally or with external partners, e.g. Google

- Accelerated Python (78 students)
- Day of TensorFlow (63 students)
- Research Computing for Social Scientists (12 students)
- Working with Social Sciences Data in R (18 students)
- Social Sciences Data in Python (16 students)
- Introduction to TensorNetwork (17 students)
Events in FY20: Workshops

Workshops = 1.5 - 2 hour training to advance computational skills in a group setting. Range of 8-69 attendees per workshop.

- Introduction to Linux (x 2)
- Introduction to Scripting (x2)
- Introduction to High Performance Computing (x3)
- Text Analysis I: Introduction to Computational Text Analysis.
- Text Analysis II: Statistical Approaches.
- Text Analysis III: Advanced Methods
- Practical Applications of Machine Learning in Python
- Introduction to Machine Learning with scikit-learn (x2)
- Analyzing Financial Time Series Data using Python
**Python User Group** = community computational group that meets every two weeks for all students and postdocs. 144 attendees in FY20.

- Intermediate NLP with spaCy
- Pandas — The Bare Basics
- Training an Optical Character Recognition (OCR) Model
- Extracting Data from APIs
- Probabilistic Programming with Pyro
- Implementing Historical Algorithms
- Image Classification with PyTorch
A&S and SEAS Students by Population FY20

A&S total students = 142
- PhD 74
- Postdocs 15
- Masters 53

SEAS total students = 68
- PhD 25
- Postdocs 11
- Masters 32

Percentages by position:
- PhD 47%
- Postdocs 12%
- Masters 40%
<table>
<thead>
<tr>
<th>Year 1 total students bootcamps = 235</th>
<th>Year 3 goal for bootcamps and intensives = 600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 2 students &amp; postdocs bootcamps &amp; intensives = 508</td>
<td>Year 3 goal for all training events = 900</td>
</tr>
<tr>
<td>Year 2 graduate students &amp; postdocs in all training events = 844</td>
<td>Year 1 &amp; 2 A&amp;S students trained = 187</td>
</tr>
<tr>
<td>Year 2 total events = 36 (to date)</td>
<td>Year 1 &amp; 2 SEAS students trained = 123</td>
</tr>
<tr>
<td>Current number of SC instructors = 35</td>
<td>Year 2 postdocs trained = 84</td>
</tr>
</tbody>
</table>
## 2020 Budget and Actual

<table>
<thead>
<tr>
<th>Type</th>
<th>FY20 Budget</th>
<th>FY20 Actual</th>
<th>H/L than Budget</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>$16,000</td>
<td>$15,800</td>
<td>$200</td>
</tr>
<tr>
<td>Bootcamps</td>
<td>$35,000</td>
<td>$30,583</td>
<td>$4,417</td>
</tr>
<tr>
<td>Distinguished Lectures</td>
<td>$11,300</td>
<td>$5,300</td>
<td>$6,000</td>
</tr>
<tr>
<td>Curricular Development</td>
<td>$11,000</td>
<td>$7,500</td>
<td>$3,500</td>
</tr>
<tr>
<td>General Support</td>
<td>$8,000</td>
<td>$2,909</td>
<td>$5,091</td>
</tr>
<tr>
<td>FY19 Fund balance</td>
<td></td>
<td></td>
<td>$2,011</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$81,300</strong></td>
<td><strong>$62,092</strong></td>
<td><strong>$21,219</strong></td>
</tr>
</tbody>
</table>
## FY21 Budget

<table>
<thead>
<tr>
<th>Type</th>
<th>FY20 Expenses</th>
<th>FY21 Budget</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>$15,800</td>
<td>$16,000</td>
</tr>
<tr>
<td>Bootcamps</td>
<td>$30,583</td>
<td>$35,000</td>
</tr>
<tr>
<td>Distinguished Lectures</td>
<td>$5,300</td>
<td>$6,719</td>
</tr>
<tr>
<td>Curricular Development</td>
<td>$7,500</td>
<td>$9,000</td>
</tr>
<tr>
<td>General Support</td>
<td>$2,909</td>
<td>$4,500</td>
</tr>
<tr>
<td>Program Total</td>
<td>$62,092</td>
<td>$71,219</td>
</tr>
</tbody>
</table>
## FY21 Funding Pledged

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>FY21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel</td>
<td>Library</td>
<td>$102,244</td>
</tr>
<tr>
<td>Program</td>
<td>SEAS</td>
<td>$25,000</td>
</tr>
<tr>
<td>Program</td>
<td>A&amp;S</td>
<td>$25,000</td>
</tr>
<tr>
<td>Program</td>
<td>FY20 unspent</td>
<td>$21,219</td>
</tr>
<tr>
<td>Program Total</td>
<td></td>
<td><strong>$71,219</strong></td>
</tr>
</tbody>
</table>