Shared Research Computing Policy Advisory Committee

Fall 2019 Meeting

Tuesday, December 3, 2019

10:00 a.m. – 11:30 a.m.

COLUMBIA | RESEARCH

Shared Research Computing
Policy Advisory Committee

Agenda

- Introductions
- High Performance Computing Update
- Foundations Update
- Research Data Platform
- Other Business

COLUMBIA | RESEARCH

Shared Research Computing
Policy Advisory Committee

Introductions

Chris Marianetti, Chair of SRCPAC

COLUMBIA | RESEARCH

Shared Research Computing
Policy Advisory Committee

High Performance Computing Update

Kyle Mandli, Chair of the HPC Operating Committee

George Garrett, Manager of Research Computing

COLUMBIA | RESEARCH

Shared Research Computing
Policy Advisory Committee

HPC Agenda

- Governance
- Support
- Cluster Updates, Stats, Expansion
  - Terremoto
  - Habanero
- Data Center Cooling Expansion Update
- Upcoming Cluster Update

HPC Governance

- Shared HPC is governed by the faculty-led HPC Operating Committee, chaired by **Kyle Mandli**.
- The committee reviews business and usage rules in open, semiannual meetings.
- The last meeting was held on November 12, 2019. Next meeting will be in Spring 2020.
- **All HPC Users (Terremoto, Habanero) are invited.**

HPC Support Services

- **Email**
  - hpc-support@columbia.edu
- **Office Hours**
  - In-person support from 3pm – 5pm on 1st Monday of month
  - RSVP required (Science & Engineering Library, NWC Building)
- **Group Information Sessions**
  - HPC support staff meet with your group
- **Training Workshops every semester**
  - Introduction to Linux
  - Introduction to Scripting
  - Introduction to High Performance Computing

Cloud Computing Consulting

- Overview of features of cloud service providers (AWS, Google, Azure)
- Cost estimates and planning workflow for efficiency and price
- Creation and initial configuration of images, including software installation

Terremoto

TERREMOTO

- Launched in December 2018
- Expanding in December 2019

Terremoto - Expansion

- **27 Compute Nodes (648 cores)**
  - 19 Standard Nodes
  - 4 High Memory nodes
  - 4 GPU nodes with NVIDIA V100 GPUs
- **80 TB of storage of additional storage**
- **10 Research Groups participated**

Terremoto Expansion machines have arrived and are going live December 2019.

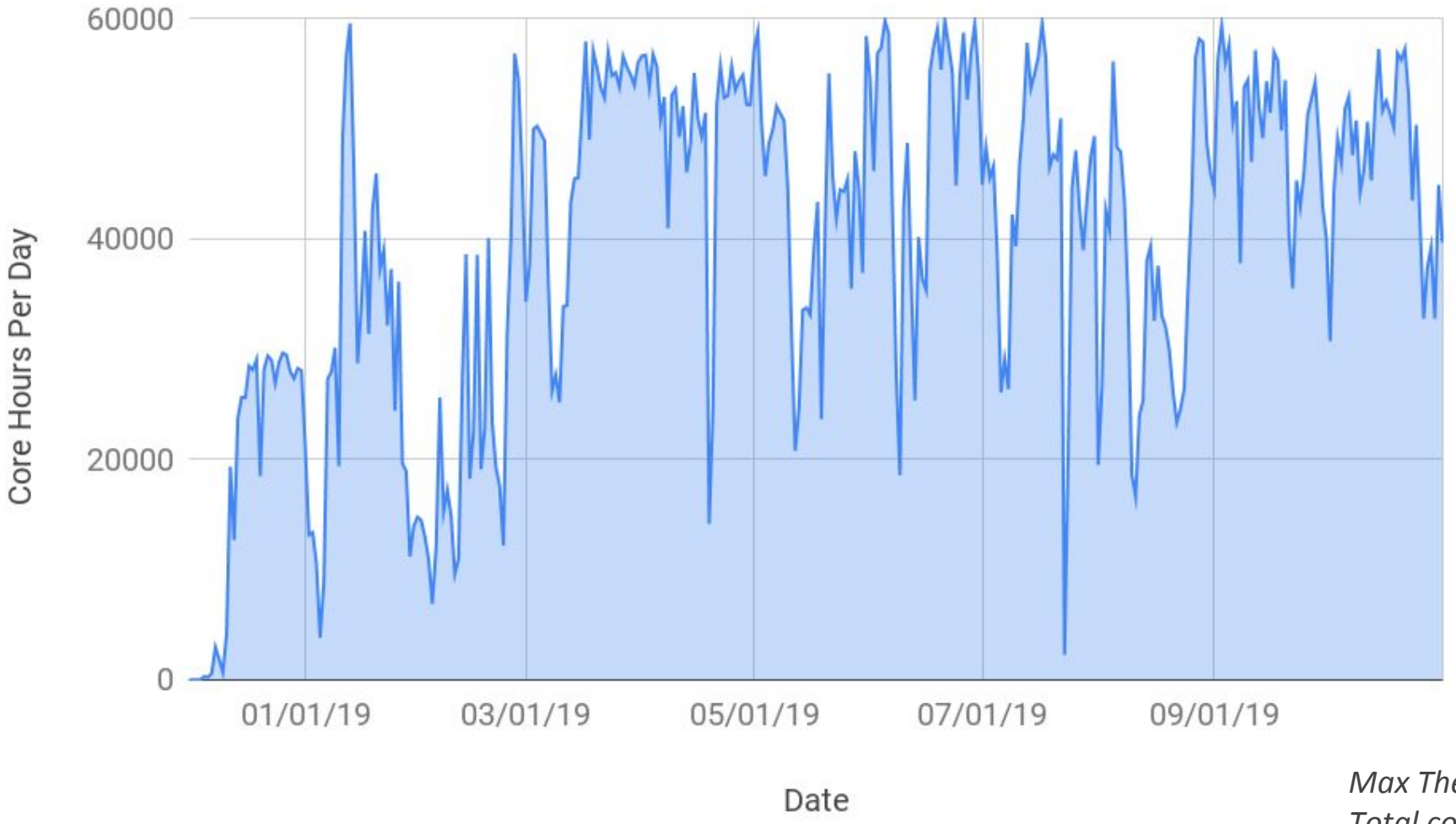
Terremoto Specifications - After Expansion

- **137 Compute Nodes (3,288 cores)**
  - 111 Standard nodes (192 GB)
  - 14 High Memory nodes (768 GB)
  - 12 GPU nodes with NVIDIA V100 GPUs
- **500 TB storage** (Data Direct Networks GPFS GS7K)
- **Dual Skylake Gold 6126 cpus**, 2.6 Ghz, AVX-512
- **100 Gb/s EDR** Infiniband
- **480 GB SSD** local drives

Terremoto - Participation and Usage

- **29** research groups
- **325** users
- **Over 14 million core hours** utilized this year
- **5 year** lifetime

Terremoto - Cluster Usage in Core Hours



*Max Theoretical Core Hours Per Day = **63,360***
*Total core hours used since launch: **14 million***

Habanero

Habanero - Specifications

Specs

- **302 nodes** (7248 cores)
  - 234 Standard servers
  - 41 High memory servers
  - 27 GPU servers
- **800 TB** storage (DDN GS7K GPFS)

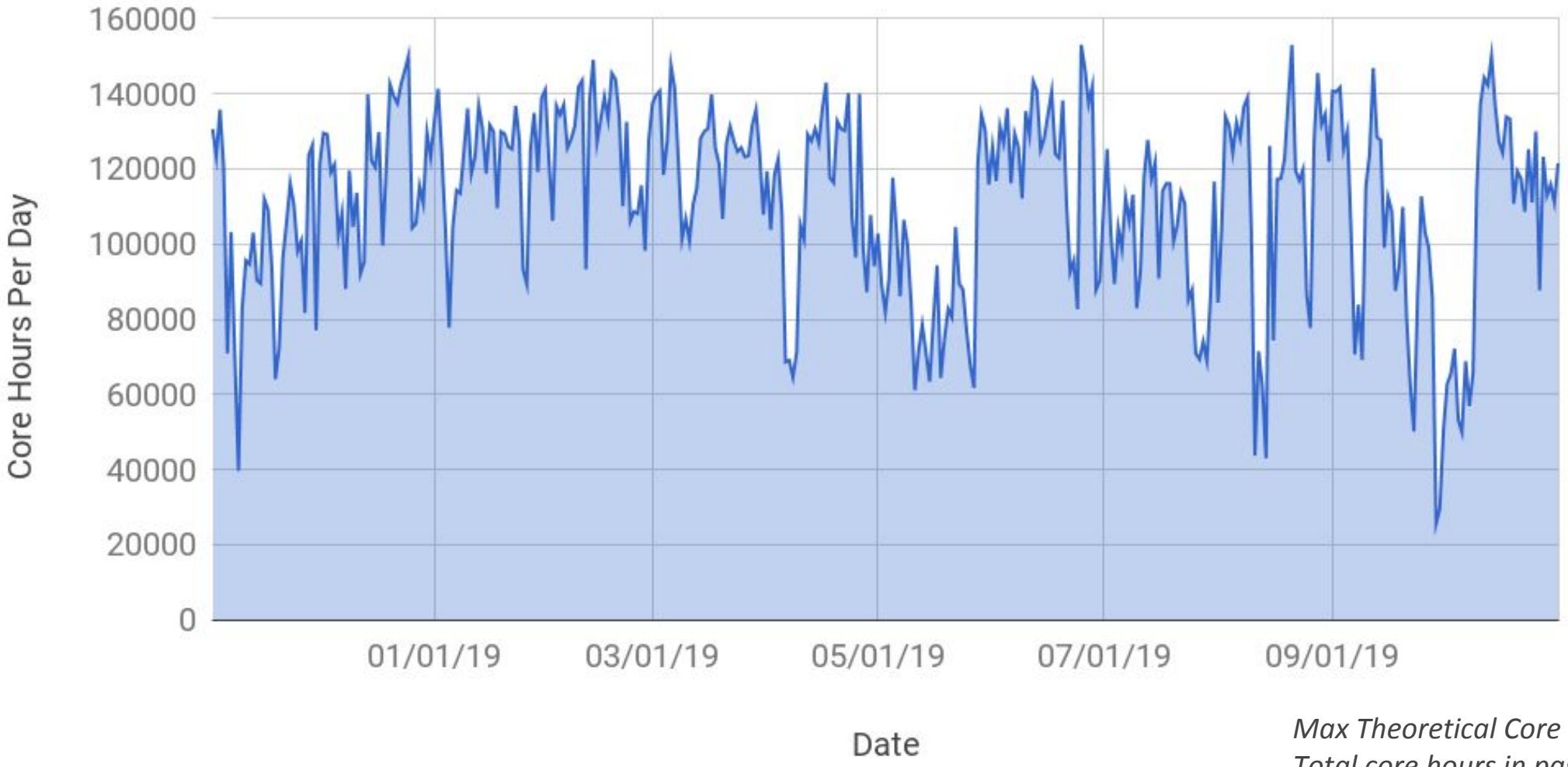
Lifespan

- *222 nodes expire December 2020*
- *80 nodes expire December 2021*

Habanero - Participation and Usage

- **44** groups
- **1,897** users since launch (**215** active)
  - (**347** added since Spring 2019)
- **20** renters since launch
- **317** free tier users since launch (**33** active)
- Education tier
  - **18 courses** since launch (**3** added since Spring 2019)

Habanero - Cluster Usage in Core Hours



*Max Theoretical Core Hours Per Day = **174,528***
*Total core hours in past 12 months: **41 million***

HPC Updates - Singularity and Open OnDemand

Singularity

- Easy to use, Docker-like containers for HPC
- Enables reproducibility and simplifies software deployment
- Bring your own container (use on Laptop, HPC, etc.)
- Available now on both Terremoto and Habanero

Open OnDemand HPC Web Portal

- Interactive HPC via your web browser.
- Get an SSH shell, submit and monitor jobs through your browser.
- https://openondemand.org
- ***Piloting on Terremoto***. Contact us if interested in trying it out.

OPEN OnDemand

Data Center Cooling Expansion Complete

- Data Center cooling expansion project was completed in July 2019
- A&S, SEAS, EVPR, and CUIT contributed to expand Data Center cooling capacity
- Assures HPC capacity for next several generations

Upcoming Cluster - Spring 2020

- In planning stages, details depend on demand
- Announcement of buy-in opportunity typically sent in April
- New machine types likely, RFP possible
- Purchase round would commence late Spring 2020
- Go-live 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

Foundations for Research Computing Update

Marc Spiegelman, Chair of the Advisory Committee

Barbara Rockenbach, Associate University Librarian for Research and Learning

COLUMBIA|RESEARCH

Shared Research Computing
Policy Advisory Committee

2019/2020 Academic Year Target Goal

Target goal for 2019/20 academic year to reach 500 students through novice boot camps, intermediate intensives, and workshops

Novice Boot Camps = 124

Intermediate Intensives = 187

Workshops = 151

Total = 462 Students

Novice boot camps = 2 day training based on Software Carpentry curriculum for novice learners

Intermediate intensives = 1 day training for intermediate learners with curriculum developed internally or with external partners e.g. Google

Workshops = 1.5 - 2 hour training opportunity to advance computational skills in a group setting

Events since August, 2019

BOOTCAMPS & INTENSIVES

- Accelerated Python
- Day of TensorFlow
- Intro to Research Computing Bootcamp
- Research Computing for Social Scientists
- Working with Social Sciences Data in R
- Social Sciences Data in Python

WORKSHOPS

- Introduction to Linux
- Introduction to Scripting
- Introduction to High Performance Computing
- 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

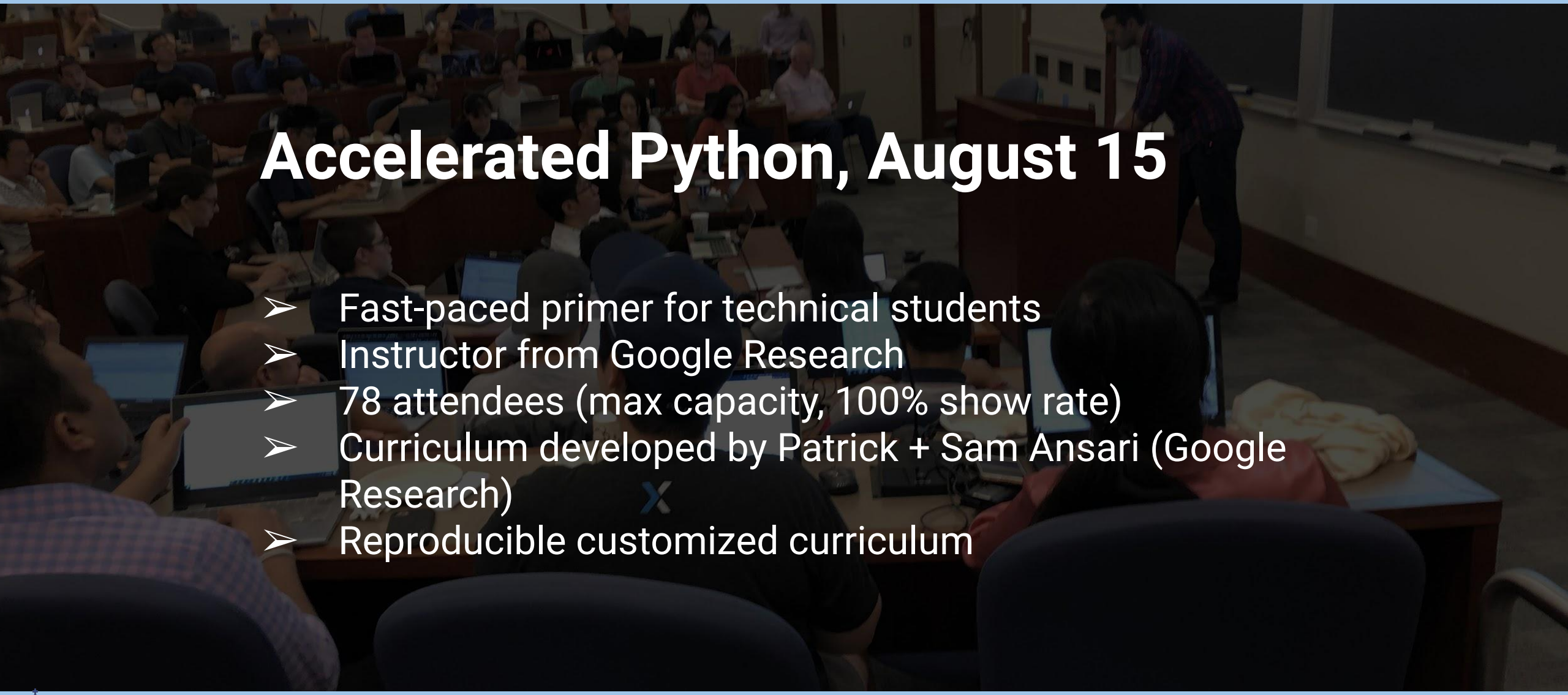
PYTHON USER GROUP

- 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

Distinguished Lecture: Stephanie Hankey

Novice Bootcamp, August 26-27

- Taught Unix, Git, and Python or R
- Software Carpentry curriculum
- 124 attendees (max capacity, 100% show rate)



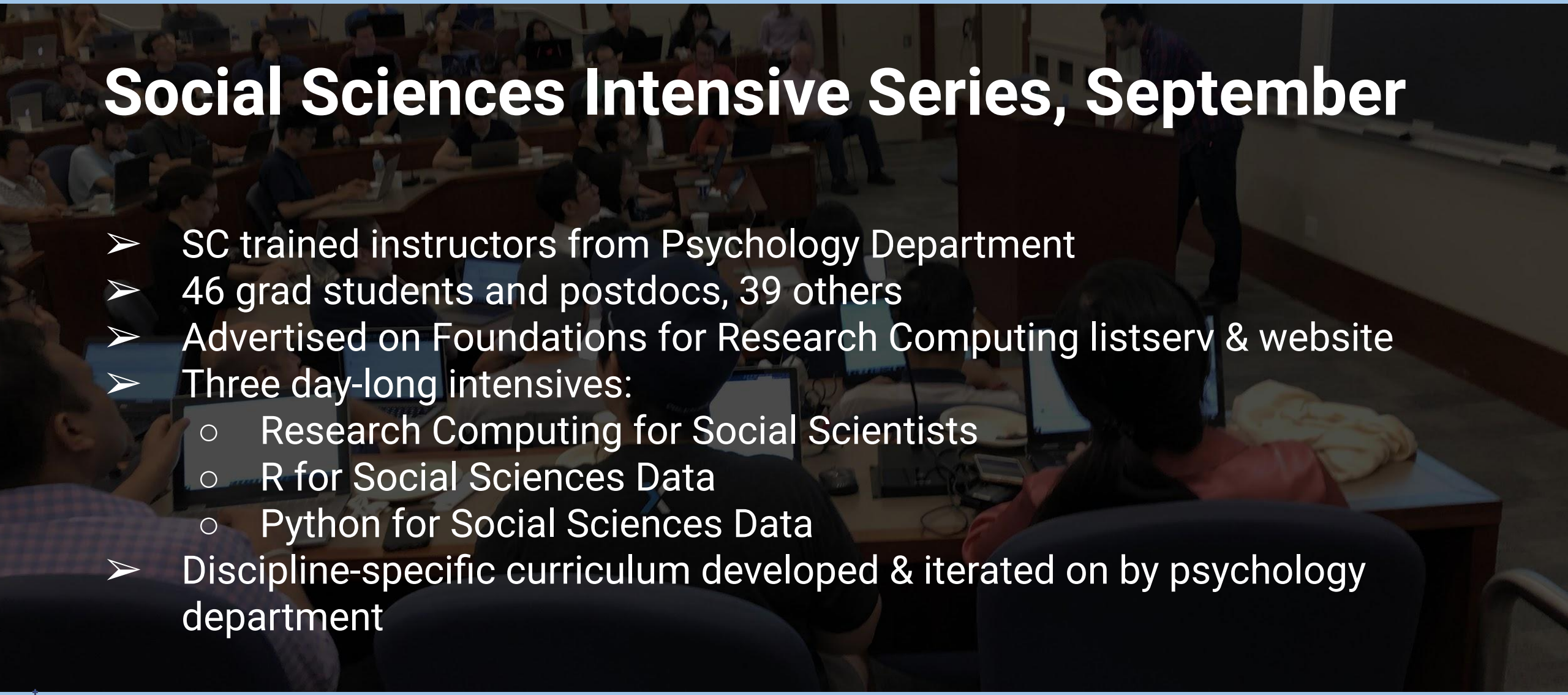
Accelerated Python, August 15

- Fast-paced primer for technical students
- Instructor from Google Research
- 78 attendees (max capacity, 100% show rate)
- Curriculum developed by Patrick + Sam Ansari (Google Research)
- Reproducible customized curriculum

Bootcamps and Intensives

TensorFlow, August 16

- Taught new TensorFlow 2.0
- Instructor from TensorFlow team
- 63 attendees (max capacity, 100% show rate)



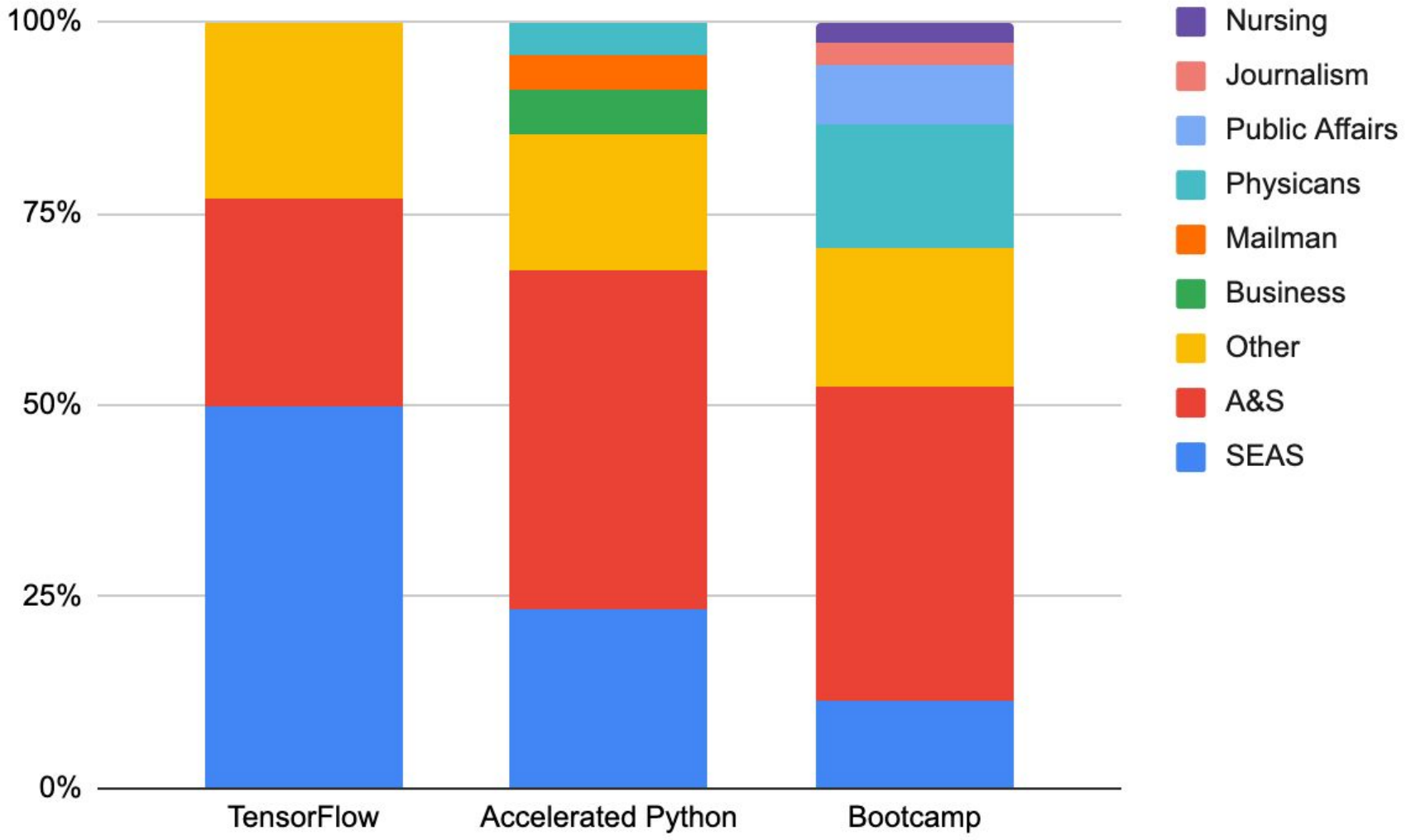
Social Sciences Intensive Series, September

- SC trained 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

What they're saying...

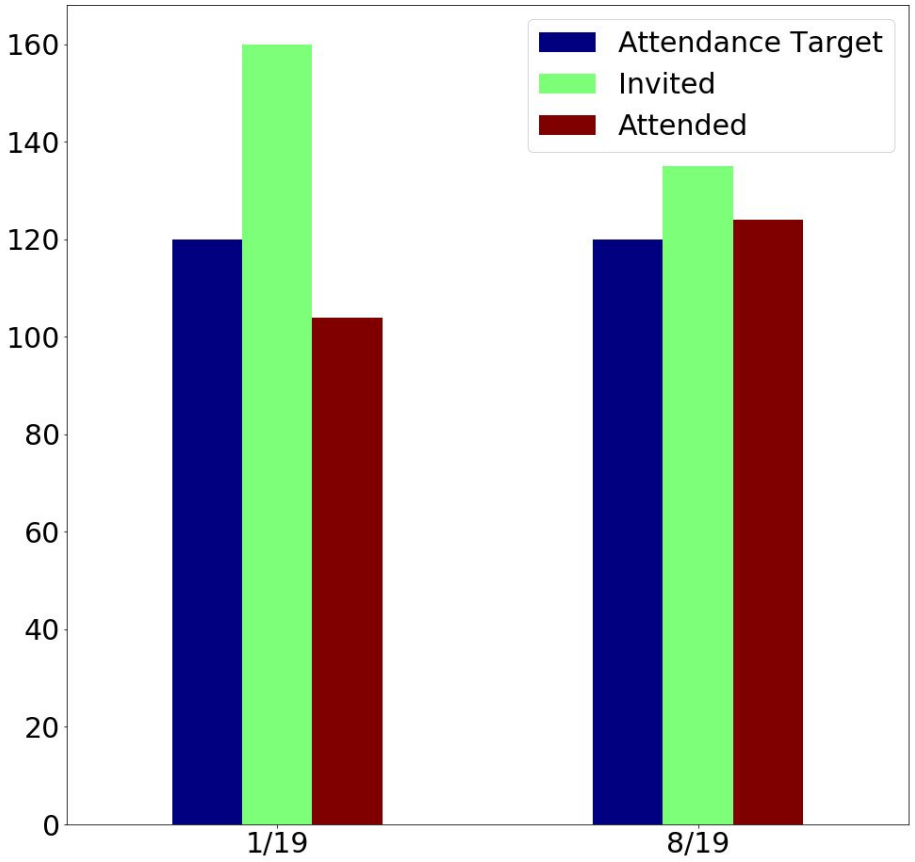
Accelerated Python	I was happily amazed that we could cover in one single session from the very basics until clustering and prediction models in a very smoothly and versatile manner.
TensorFlow	The content of this event is extremely helpful for someone that has a bit experience in machine learning but wish to have a taste of deep learning and TensorFlow.
R for Social Sciences	The instructor was very good. I liked the interactive questions. She was very helpful going step by step to new beginners like me. Did not just read off an outline.

August Demographics



Application Sorting Process Implementation

Invitation Process Comparison

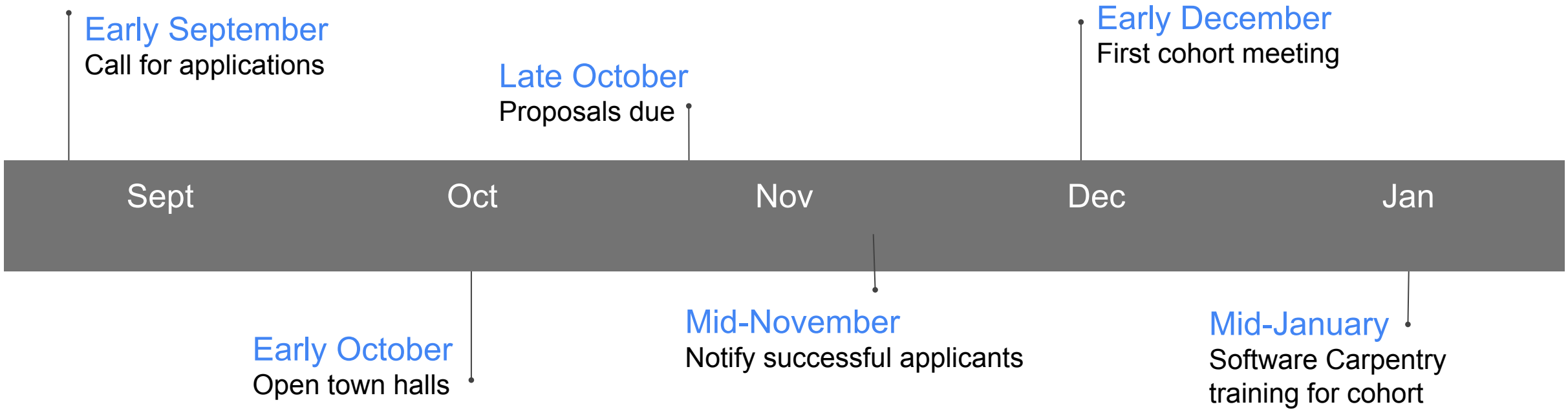


- 420 completed applications eligible for bootcamps and intensives
- Received fewer but higher-quality applications
- New application filtering process
- New RSVP process
- Comparing January to August, no-show rate reduced from 35% to 8%

Curricular Innovation Grant (CIG)

CIG Fellows are grad students and postdocs who create curriculum modules for RC Foundations programming

22 applications received for six slots (increased to eight with support from QMSS)



CIG Proposal Topics Selected By Committee

1. Interactive Data Visualization with R & Shiny
2. Intro to Deep Learning with PyTorch
3. Wrangling Multilevel Data with R & the Tidyverse
4. Data Analysis and Manipulation with Xarray
5. Python for the Analysis and Visualization of Biological Datasets

With support from Quantitative Methods in the Social Sciences:

6. Tidying Survey Data in R
7. Data Visualization in R (ggplot2)

Questions for Discussion

- What emerging tools and methods should we be following, e.g. TensorFlow?
- How do we better integrate novice training and intermediate training with the needs of departments?
- What are new audiences for boot camps? (undergrads, faculty)?

Research Data Platform

Maneesha Aggarwal, AVP, Academic and Research Services

COLUMBIA | RESEARCH

Shared Research Computing
Policy Advisory Committee

What is a Data Platform

- An environment to:
  - Store, organize, share
  - Analyze / Visualize
  - Web based publications
  - Archive data

Origin

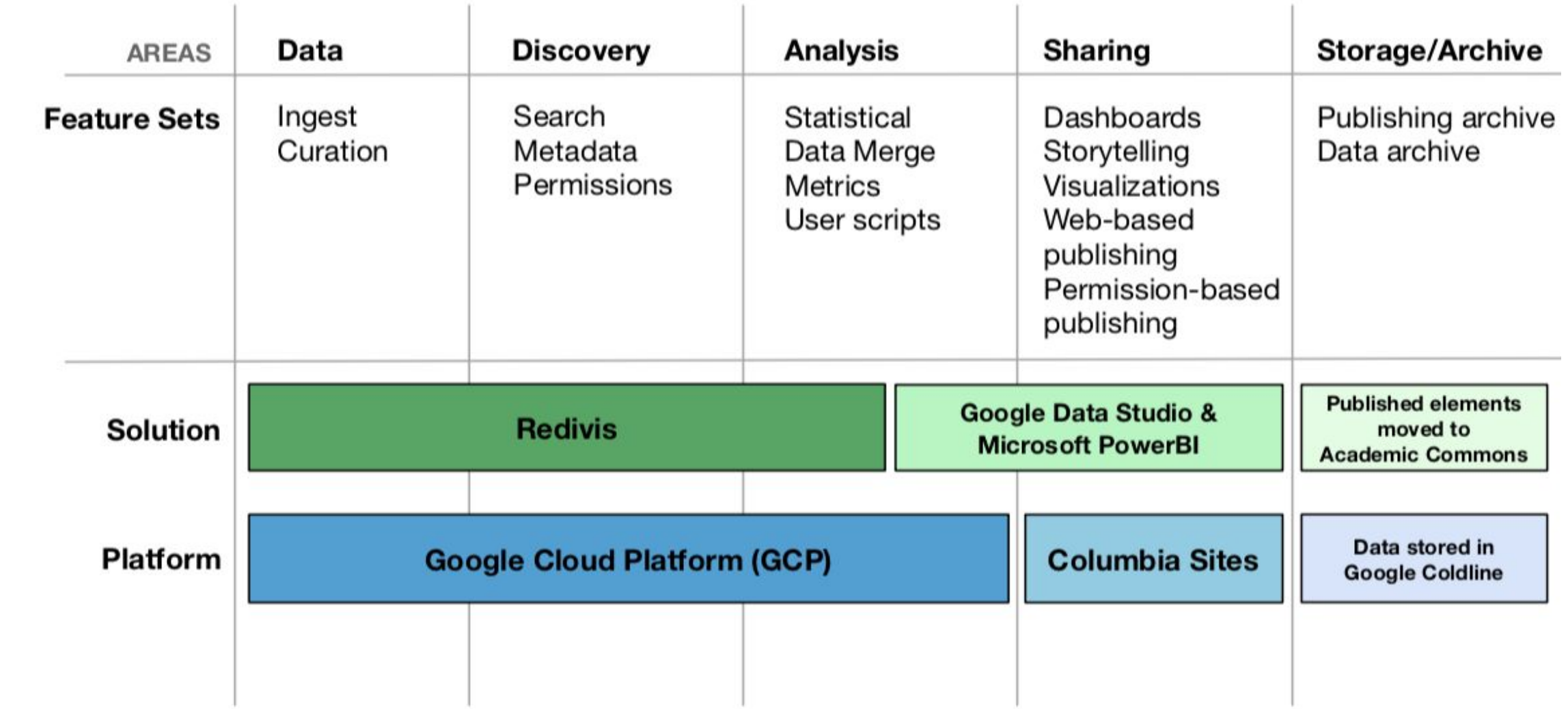
- Summer: 2018 Columbia World Projects
  - Share data for the greater good
  - Different projects, data, different environments

- Examples
  - Climate Data - Lisa Goddard
  - Energy Data - Prof Vijay Modi
  - Ocean Data - Ryan Abernathy

Concept: Academic Cloud

- Connect disparate datasets across disciplines
- Allow users easy access to these datasets for analysis
- Options for data simulation, analysis, modelling, visualization
- Secure, Scalable, and Sustainable

Architectural Approach

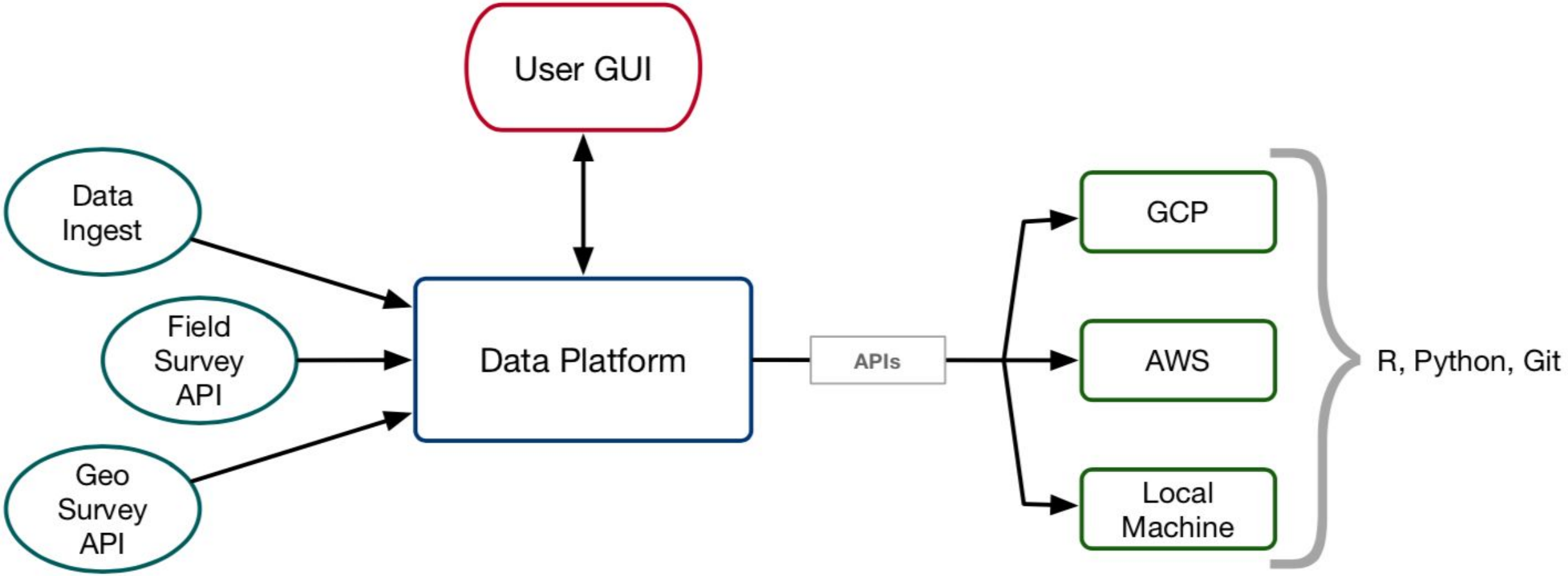


Platform Capabilities

- Upload, store, curate
  - Data types: tabular, shape, images
- Share and Discover/ Collaborate
  - Merge data / data mashups / create new datasets
  - Reproducibility
- Analyze / Visualize
  - Research Work Bench
- Web based publication
- Archive data

- Pay for what you need

Access to Platform



Demo

Dataplatform.cuit.columbia.edu

Engage / Contact

dataplatform-admin@columbia.edu

Maneesha@columbia.edu

Other Business

Chris Marianetti, Chair of SRCPAC

CloudBank

- Front line user support, cloud solution consulting, training, and assistance in preparing proposals that include cloud resources.
- Through aggregating multiple small requests and innovative financial contract types, CloudBank will pass along savings and more flexible terms to researchers that would otherwise be unavailable to them.
- Initially provide access to Amazon AWS, Google GCP & Microsoft Azure.
- NSF will control allocations as part of awarding proposals.
- Cloud allocations will not bear indirect costs.

Email : srcpac@Columbia.edu
Website:
https://research.columbia.edu/content/srcpac

Thank you