Shared Research Computing Policy Advisory Committee

Fall 2018 Meeting
Friday, December 7th

Welcome and Introductions

Chris Marianetti
Chair of SRCPAC

Today's Agenda

Welcome and Introductions
Chris Marianetti

HPC Update
Kyle Mandli and George Garrett

***Foundations for Research Computing* Update**
Marc Spiegelman and Barbara Rockenbach

Research Data Survey
Barbara Rockenbach

Potential for New Subcommittees
Chris Marianetti

CUIT Updates
Research Computing Services Group

Closing Remarks
Chris Marianetti

HPC Update

Kyle Mandli
Chair of Operating Committee

George Garrett
Manager of Research Computing Services

High Performance Computing Update

Topics

- Governance
- Support
- Yeti
- Habanero
- Terremoto
- Data Center Cooling Expansion 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.

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 present with your group
- Topics can be general/introductory or tailored
- Contact hpc-support@columbia.edu to schedule an appointment

Yeti Cluster Update

- **Yeti Round 1** retired November 2017
- **Yeti Round 2** to retire March 2019

Habanero

Specifications

- 302 compute nodes (7,248 cores)
- 740 TB storage (DDN GS7K GPFS)
- 397 TFLOPS of processing power

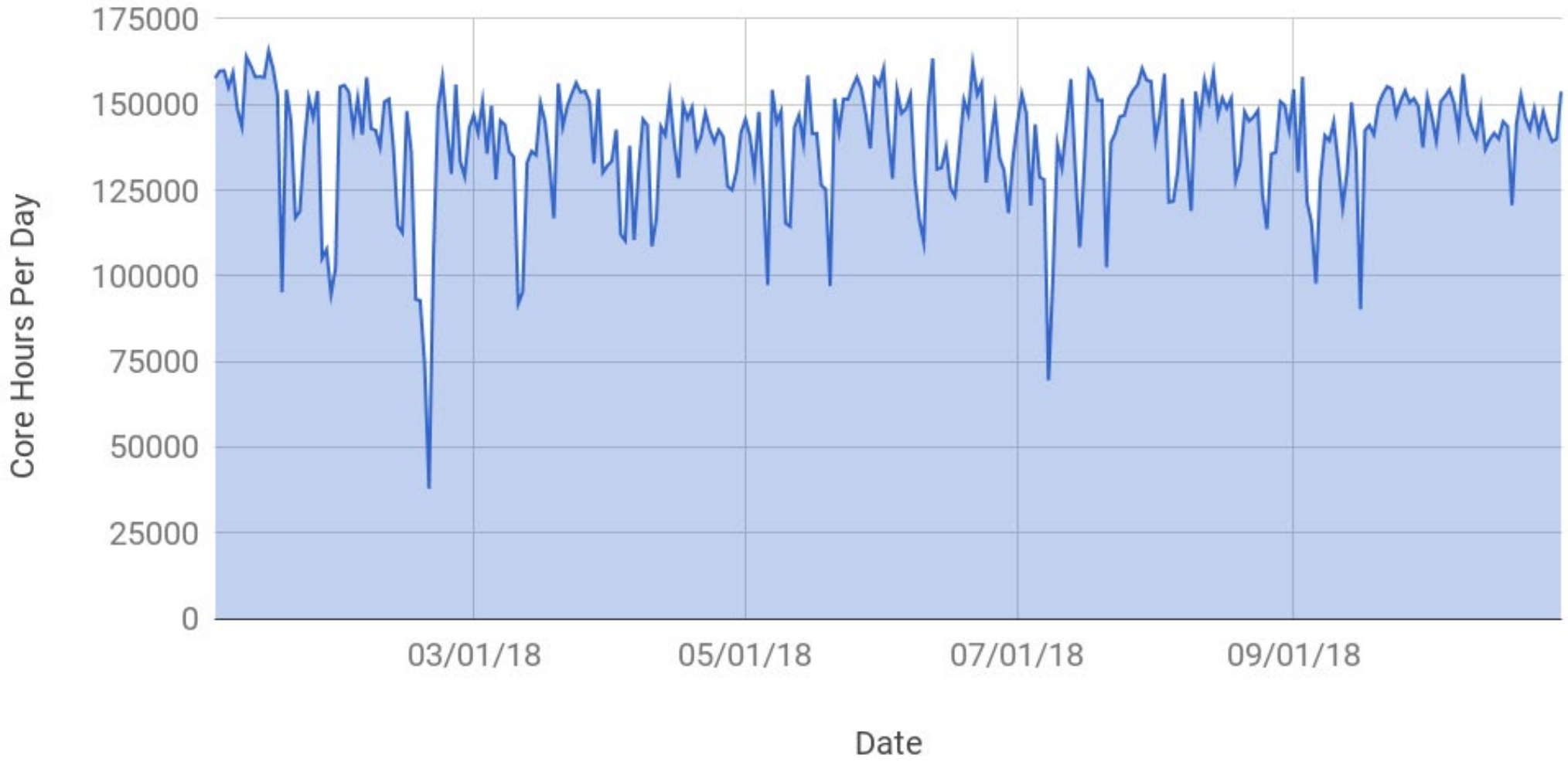
Lifespan

- 222 nodes expire 2020
- 80 nodes expire 2021

Habanero – Participation and Usage

- 44 groups
- 1,400 users
- 9 renters
- 120 free tier users
- Education tier
  - 13 courses since launch
- 2.7 million jobs completed

Habanero – Cluster Usage in Core Hours



TERREMOTO

LIVE!!

Wednesday, December 5!

- **24** research groups
- **5** year lifetime

Specifications

- **110 Compute Nodes** (2640 cores)
  - 92 Standard nodes (192 GB)
  - 10 High Memory nodes (768 GB)
  - 8 GPU nodes with 2 x NVIDIA V100 GPUs
- **430 TB storage** (Data Direct Networks GPFS GS7K)
- **255 TFLOPS** of processing power
- Dell Hardware, Dual Skylake Gold 6126 CPUs, 2.6 Ghz, AVX-512
- 100 Gb/s EDR Infiniband, 480 GB SSD drives

Data Center Cooling Expansion Update

- A&S, SEAS, EVPR, and CUIT contributed to **expand Data Center cooling capacity**
- Work to be **completed by February 2019**
- **Assures HPC capacity for several generations**

Foundations for Research Computing Update

Marc Spiegelman
Chair, *Foundations* Advisory Committee

Barbara Rockenbach
Associate University Librarian for Research and Learning

Foundation's Goals

1. Address current **needs**, and **demand** for informal training in computational science to improve research capabilities
2. Provide a hierarchical training infrastructure to serve **novice**, **intermediate**, and **advanced** users
3. Develop and foster a Columbia-wide **culture** and **community** of research computing
4. Leverage **existing University- and school-based investments** in research computing infrastructure

Hierarchical Program Structure

- **Novice**
  - Institutional Membership with ***The Carpentries*** (Software, Data, Library Carpentry)
  - Pre-semester Boot Camps: ~200+ students per year
  - Refresher Monthly Workshops & Help Room Office hours
- **Intermediate**
  - Help Room Office Hours
  - Distinguished Lectures in Computational Innovation
  - Research Symposium
  - Monthly Workshops: Discipline-specific, use of advanced libraries
- **Advanced**
  - Coordination with departmental curriculum

Current Status

COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

The Carpentries and CU Instructors

Silver membership with ***The Carpentries*** established July 2019

Instructors trained from CUIT, Libraries, Computer Science, and Business:

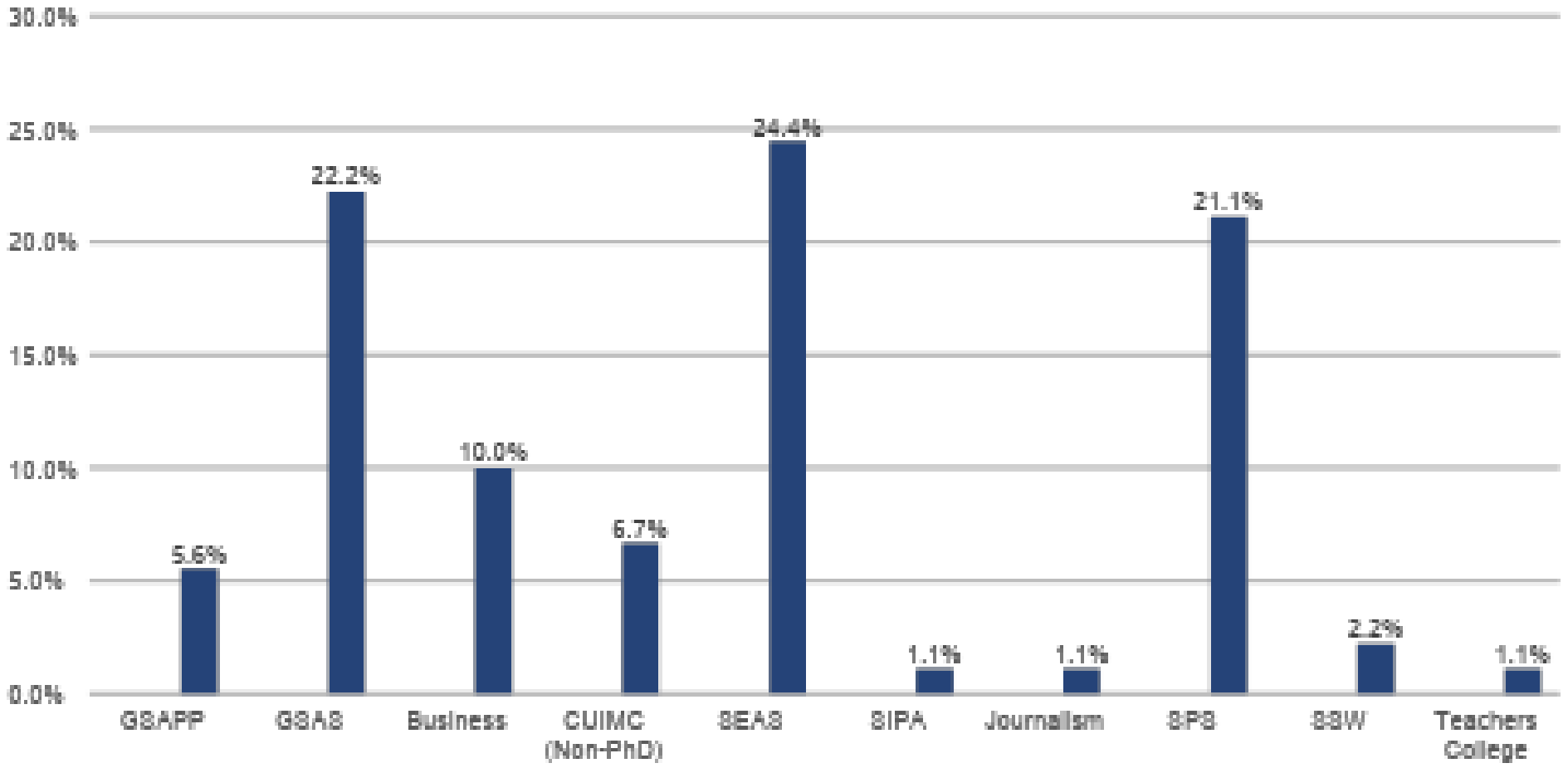
- 6 in July 2018
- 6 in October 2018

Fall 2018 Boot Camps

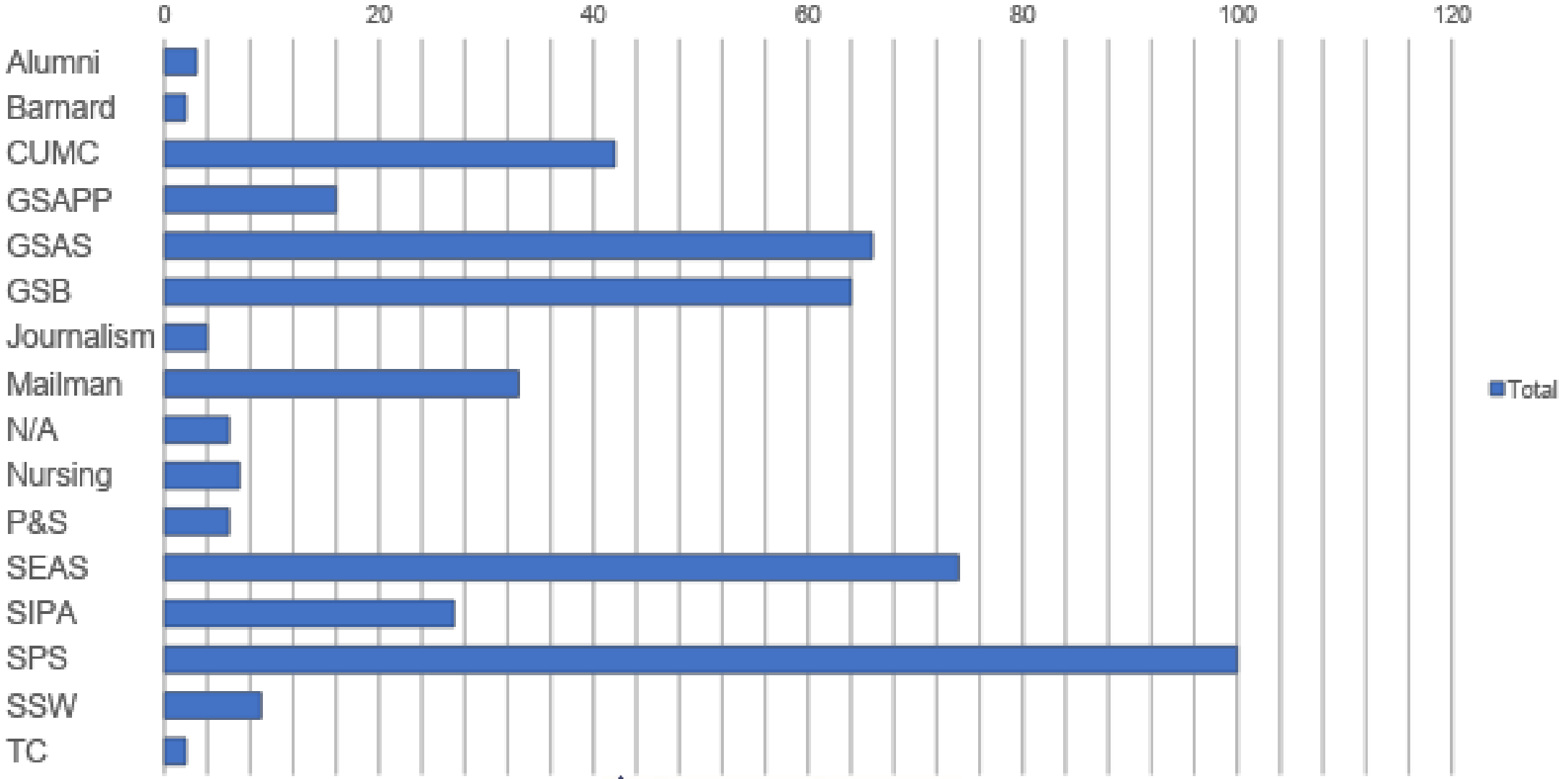
- August 27-28, 2018
- 462 registrations for 90 seats
- **90 seats filled in 4 Minutes**
- 6 instructors from CUIT, Libraries, APAM
- 3 courses from the Software Carpentries
  - Programming in **Python** (x2)
  - **R** for Reproducible Scientific Analysis

COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

Fall Bootcamp Attendance



Registration & Waitlist by School



Distinguished Lectures in Computational Innovation

All events held in Brown Institute for Media Innovation (Journalism School)

- September 13: **Bjarne Stroustrup** (creator of C++)
  - Registrations: **200** Attendance: 100+ **(Standing Room Only)**

- October 11: **Lorena Barba** (Reproducible Science and Open-Source Initiative)
  - Registrations: 90 Attendance: 40+

- November 8: **Eric Xing** (leader in commercialization of machine learning technologies)
  - Registrations: 191 Attendance: 80+

Workshops Held

Three-Part ***Introduction to HPC*** Series

(RCS held at Science and Engineering Library)

Additional workshops from Libraries:

- Panel and Survey Data Analysis Using Stata (2 sessions)
- Introduction to Data Visualization in R

Office Hours

Four Graduate Students Staffing Two Locations Each Week:

- **Mondays 3–5pm**, Science and Engineering Library
- **Fridays 1–3pm**, Butler Library

Uptake has been slow:

- Will revisit allocation of student assistants for the spring

Recruiting Program Coordinator

- Interviews conducted in several rounds throughout Fall semester
- Input from Advisory and Coordinating Committees
- **Currently finalizing offer to top candidate**

Spring Look-Ahead: Boot Camps

- January 17-18, 2019: Butler Library
- Applications open December 11
- Expanding to 4 boot camps
  - Possibility of training additional instructors
- Experienced and novice instructors paired
  - CTL Microteaching Sessions in Arrangement
- Python groups will pilot new(er) curriculum in **Plotting and Programming**
- R group will use **R for Social Scientists** module

Spring Lectures Schedule

- February 14: **Krishna Ratakonda** (IBM Fellow & CTO, Blockchain Solutions)
- March 14: **Runa Sandvik** (computer security and encryption expert)
- April 11: **Gina Helfrich** (communications and diversity initiatives)
- May 9: **Fernando Perez** (creator of iPython computer environment)

Future Plans and Opportunities

- Increase and Improve intermediate-level offerings: Need input and feedback from instructors/departments/students on most-needed content. **Huge role for coordinator.**
- Consider mechanisms for potential **curriculum development** (e.g. NRT/Carpentries, seed funding).
- **Understand demand and scale** to meet it while maintaining quality.
- Already some future discussions with additional units (CUIMC, SPS, etc).
- **All input greatly appreciated.**

Research Data Survey

Barbara Rockenbach
Associate University Librarian for Research and Learning

Potential New Subcommittees

Chris Marianetti
Chair of SRCPAC

Cloud Subcommittee

- A new subcommittee to determine how to make decisions such as:
  - When should a resident cluster **burst to a Cloud resource**?
  - How would **priorities** for use of resident vs. Cloud-based be established?

- SRCPAC membership, with support from staff should understand potential financing and charging models

GPU Subcommittee

- GPU resources have **increased astronomically in price**.
- Some peer institutions and groups have set-up **low cost, consumer grade GPU clusters.**
- SRCPAC could establish a subcommittee to assess demand, risk, and support.

CUIT Updates

Michael Weisner
Research Systems Engineer,
Columbia Population Research Center

George Garrett
Manager
Research Computing Services

Jimmy Chiong
Lead Infrastructure Engineer,
Configuration Management

Secure Data Enclave (SDE)

Service

- The SDE Provides Columbia researchers with a secure, remotely accessible, virtual Windows 10 desktop environment to store and collaboratively analyze sensitive and identifiable information.

https://cuit.columbia.edu/sde

Secure Data Enclave – Usage Requirements

- Users must have a UNI and VPN access to use the SDE. (Outside collaborators may be approved for access through proper HR registration)
- Projects have a yearly cost of **$526** per project per year
- Projects must have a sponsoring faculty member and provide a "**Data Security Officer**"

Secure Data Enclave – Features

- Members get access to a 4-core 16GB RAM Windows 10 Desktop Image
- Allows for simultaneous work by project members on data
- Certified by the Columbia University Irving Medical Center Security group for HIPAA compliance
- Supports popular statistical software packages including Stata 15, R, STAN, QGIS, and more.

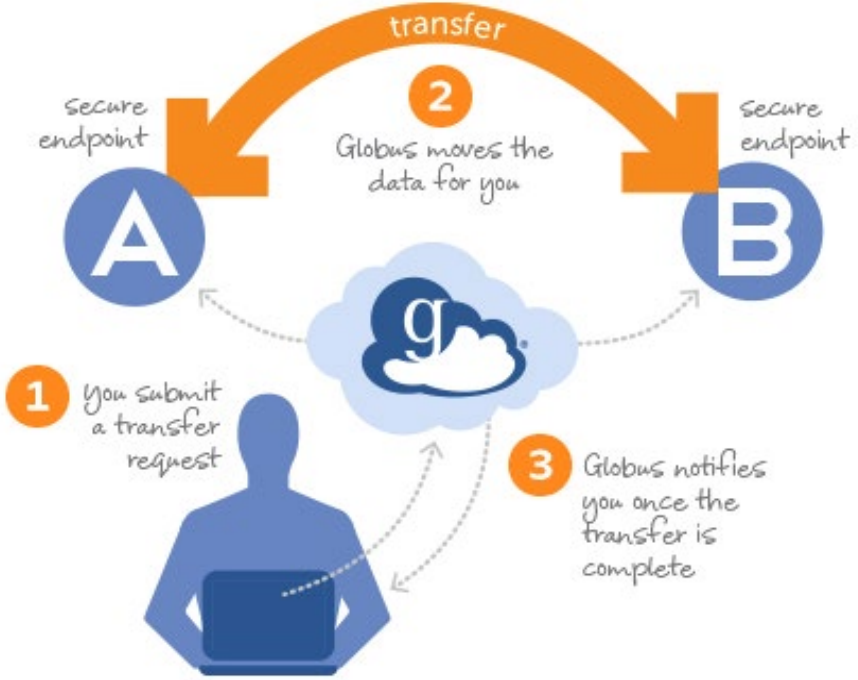
Secure Data Enclave – Data

The SDE is currently approved for use of popular datasets, including:

- The Bureau of Labor Statistics National Longitudinal Surveys (NLS) datasets
- University of North Carolina Longitudinal Study of Adolescent Health (Add Health) datasets
- European Commission Eurostat restricted economic datasets
- Department of Health records
- Restricted National Economic Data

Globus

- Provides secure, unified interface to research data.
- "***Fire and Forget***" high-performance data transfers between systems within and across organizations.
- Share data with collaborators.
- Columbia has procured an enterprise license.
- Contact **rcs@columbia.edu** to get started with Globus



AWS Enterprise Agreement

- Multiple benefits over "**click through**" agreement:

1. Improved security, privacy, and audit protections
2. Branding and intellectual property protection
3. Extended times to "exit" the service
4. Compliance with procurement and IT security policies
5. Ability to enroll in BAA (not automatic) – PHI data
6. Billing and pricing enhancements

“Linked” vs. “Delegated” Accounts

- Existing AWS accounts can “**link**” to CUIT billing family
  - Allows for central ARC billing
  - Potential to realize volume discounts over time
  - Ensures compliance with University Finance and IT security policies

“Linked” vs. “Delegated” Accounts

- For new requests, CUIT creates a “**delegated**” account:
  - SAML-based login with Columbia UNI
  - CUIT-managed CloudTrail log collection
  - Secure storage and management of the root credentials
- Researchers retain control of and responsibility for account

AWS Direct Connect

- Activated January 2018
- Provides dedicated 10 Gbps link to US-East-1 Region
- Allows direct routing of RFC1918 addresses
- Envisioned as primary link for clients between campus and their AWS resources, with VPN as backup
- Public VIF enabled allowing dedicated network performance to AWS IP addresses

Data Egress Waiver

- Research & Academic accounts can take advantage of Amazon's Data Egress Waiver program

- Credit for your data egress charges (network data traffic leaving AWS) of up to 15% of your total bill for a given month

More Information

- **Account Information**
  https://cuit.columbia.edu/aws

- **Cloud Computing Consulting**
  https://cuit.columbia.edu/cloud-research-computing-consulting

Closing Remarks

Chris Marianetti
Chair of SRCPAC