Shared Research Computing Policy Advisory Committee

Spring 2018 Meeting
Monday, April 16, 2018

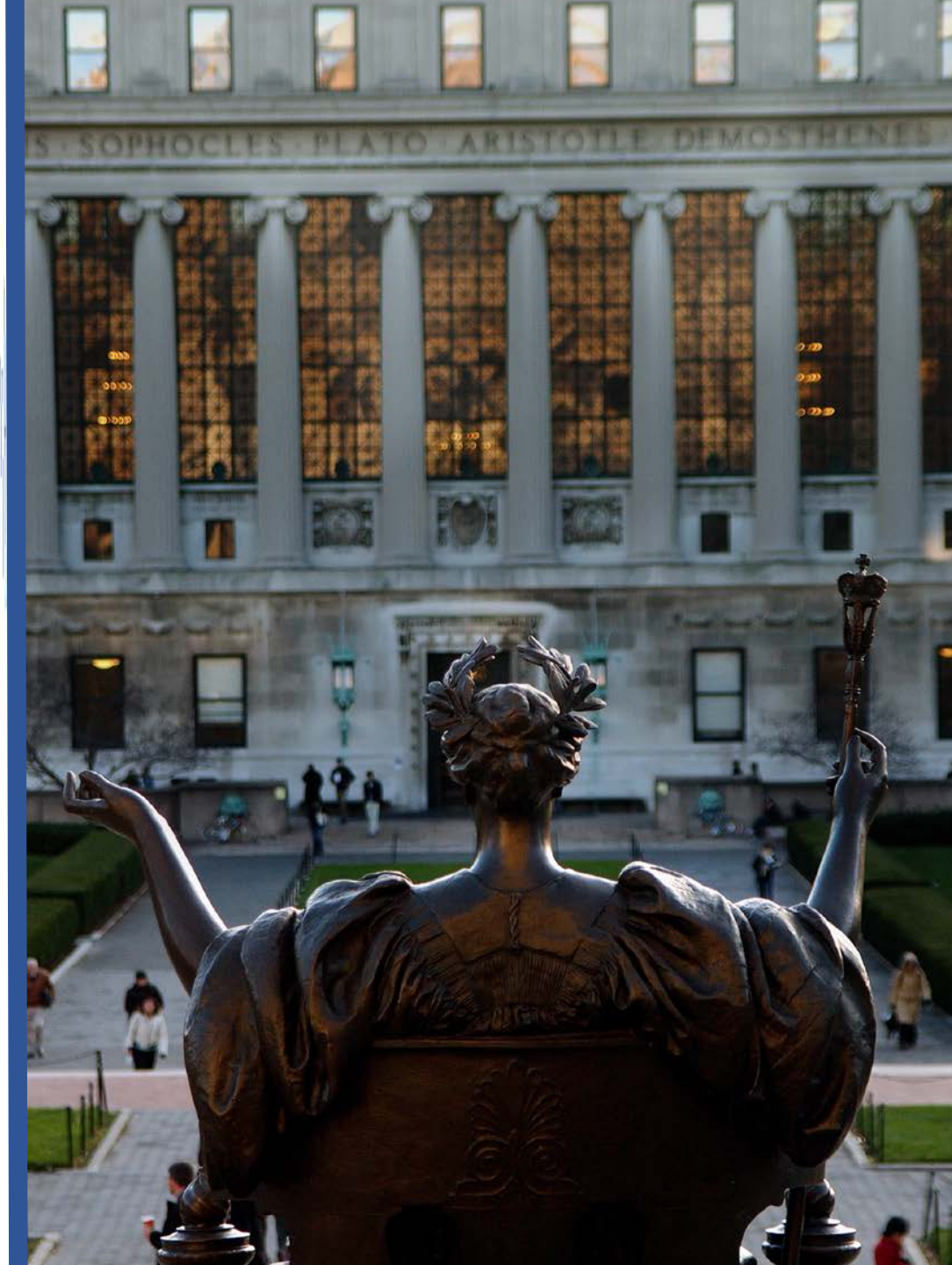
COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

Spring 2018 Agenda

Welcome & Introductions
Chris Marianetti, Chair of SRCPAC

Habanero Update
New Cluster Update
Assessing Post-Purchase Demand
Update from the Training Subcommittee
CUIT Updates
Publications Reporting

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



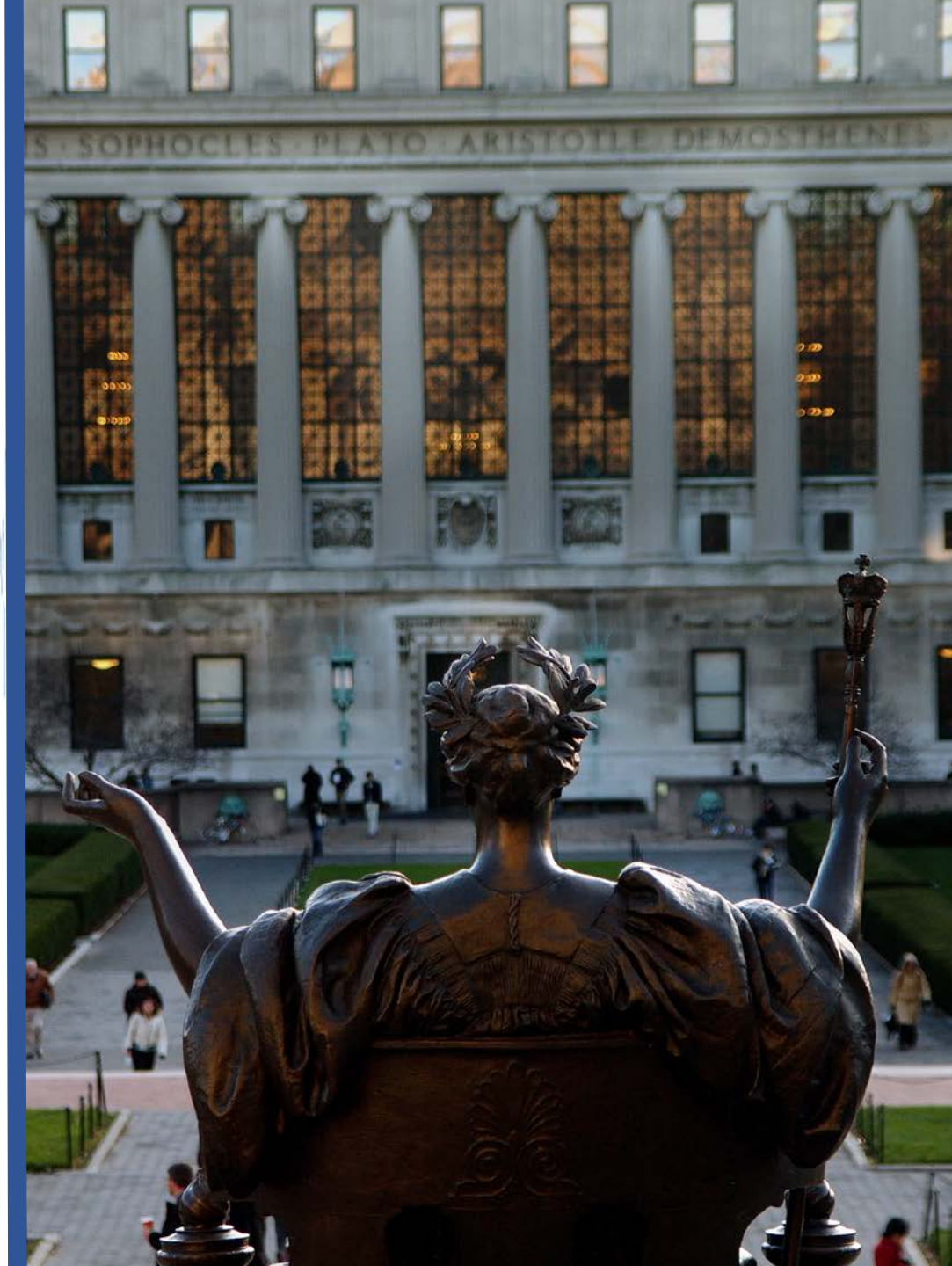
Spring 2018 Agenda

~~Welcome & Introductions~~

Habanero Update
Kyle Mandli, Chair of the Habanero Operating Committee
George Garrett, Manager of Research Computing Services

New Cluster Update
Assessing Post-Purchase Demand
Update from the Training Subcommittee
CUIT Updates
Publications Reporting

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



Our Spicy Cluster

Four Ways to Participate

1. Purchase
2. Rent
3. Free Tier
4. Education Tier

2017 Expansion Update

- 2016: 1st round launched with 222 nodes (5,328 cores)

- December 2017: Expansion nodes live
- Added 80 nodes (1,920 cores), 240 TB storage
  - 58 Standard servers (128 GB)
  - 9 High Memory servers (512 GB)
  - 13 GPU servers with 2 x Nvidia P100 modules
- 12 new research groups

- **Post-expansion total**: 302 nodes (7,248 cores)

Spring 2018 Storage Expansion

- Researchers purchased approximately 100 TB additional storage
- Order placed with vendor (DDN)
- Will install new drives upon purchasing completion
- Total Habanero storage post-expansion: 740 TB

*Contact **rcs@columbia.edu** for quota increase prior to equipment delivery.*

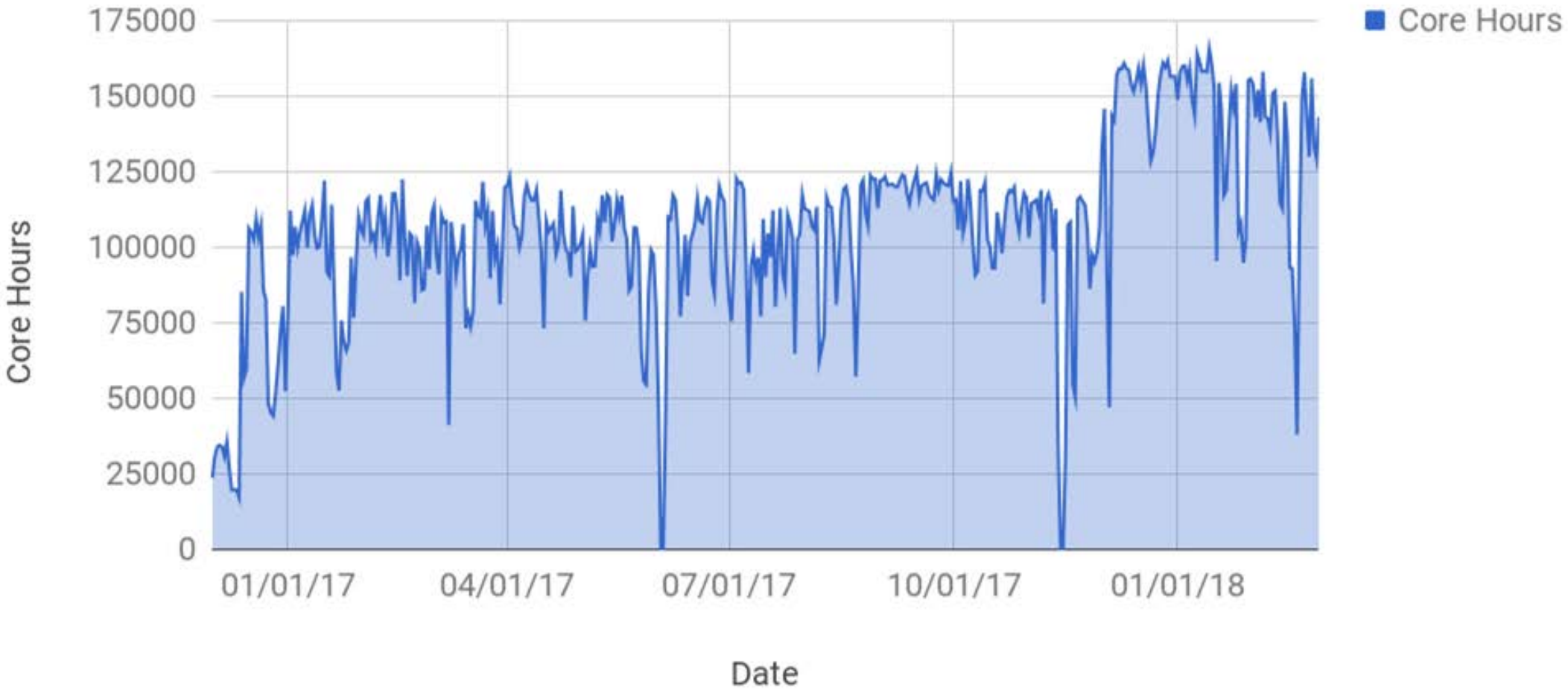
Habanero – Additional Updates

- Scheduler upgrade
  - Slurm 16.05 to 17.2
  - Bug fixes and optimizations
- New test queue added
  - High-priority short queue dedicated to interactive testing
- Jupyterhub and Docker pilot
  - Contact **rcs@columbia.edu** to participate

Habanero – Participation and Usage

- 44 groups
- 1,080 users
- 7 renters
- 63 free tier users
- Education tier
  - 9 courses since launch
  - 5 courses in Spring 2018
- 2.1 million jobs completed

Habanero – Cluster Usage in Core Hours



Habanero Business Rules

- Business rules set by Habanero Operating Committee.
- Operating Committee reviews rules in semiannual meetings.

HPC Support Services

- Email
  - **hpc-support@columbia.edu**

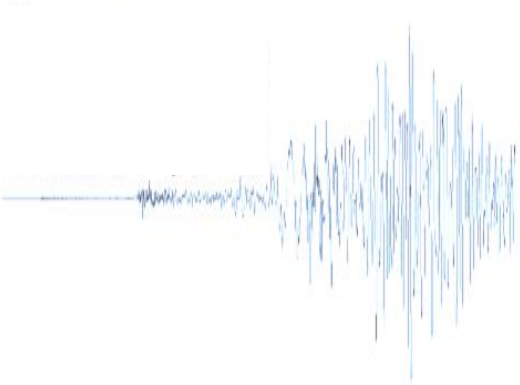
- Office Hours
  - In-person support from 3pm – 5pm on 1st Monday of month
  - RSVP required (Science & Engineering Library, NWC Building)

- 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

Workshops

Introductory workshops by CUIT & Libraries.

- Part 1: Intro to Linux
- Part 2: Intro to Scripting
- Part 3: Intro to HPC



Workshop series held in Spring and Fall.

Fall 2018 workshop schedule TBD.

HPC – Yeti Cluster Update

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

Spring 2018 Agenda

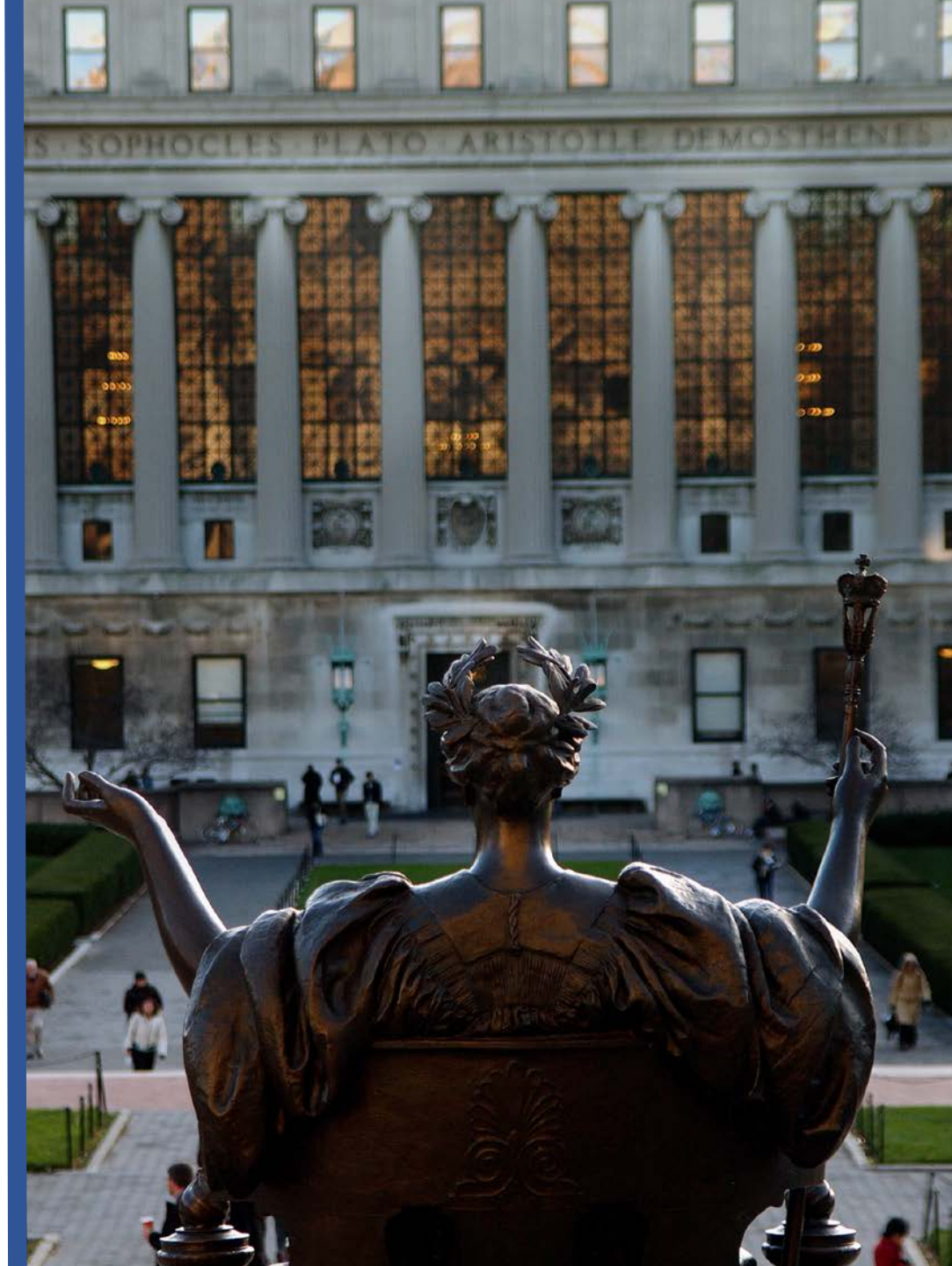
~~Welcome & Introductions~~
~~Habanero Update~~

New Cluster Update

George Garrett, Manager of Research Computing Services
Sander Antoniades, Lead Research Systems Engineer

Assessing Post-Purchase Demand
Update from the Training Subcommittee
CUIT Updates
Publications Reporting

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



8 RFP and Design Committee Members

Niko Kriegeskorte
Professor, Psychology/ZMBBI

Kyle Mandli
Assistant Professor, APAM

Bob Mawhinney
Professor and Chair, Physics

Lorenzo Sironi
Assistant Professor, Astronomy

Alan Crosswell
Chief Technologist/AVP, CUIT

Khaled Hamdy
Director, Research and Planning, Business

Rob Lane
Executive Director, IT, Computer Science

Jochen Weber
Scientific Computing Specialist, ZMBBI

New Cluster Update – Schedule

Month	Phase
February	Requirements
March	Finalize RFP
Early-April	Select Finalist Vendors
Late-April	Select Winning Vendors
May/June	Ordering
July	Finance
September	Shipping
October	Configuration & Testing
November	Production

New Cluster Update – Cooling Expansion

- A&S, SEAS, EVPR, and CUIT contributing to expand Data Center cooling capacity
- Data Center cooling expansion project has initiated
- Targeting Fall 2018 completion to house the next cluster

New Cluster – Proposed Specifications

Preliminary Menu

- Standard node (192 GB)
- High Memory node (768 GB)
- GPU node with 2 x NVIDIA V100 GPUs

All nodes will feature Dual Skylake Gold 6126 processors

- 2.6 Ghz, AVX-512, 12 cores per processor (24 total cores)

Specifications not yet finalized; subject to change from pricing/other factors.

Name Selected by RFP Committee

Spring 2018 Agenda

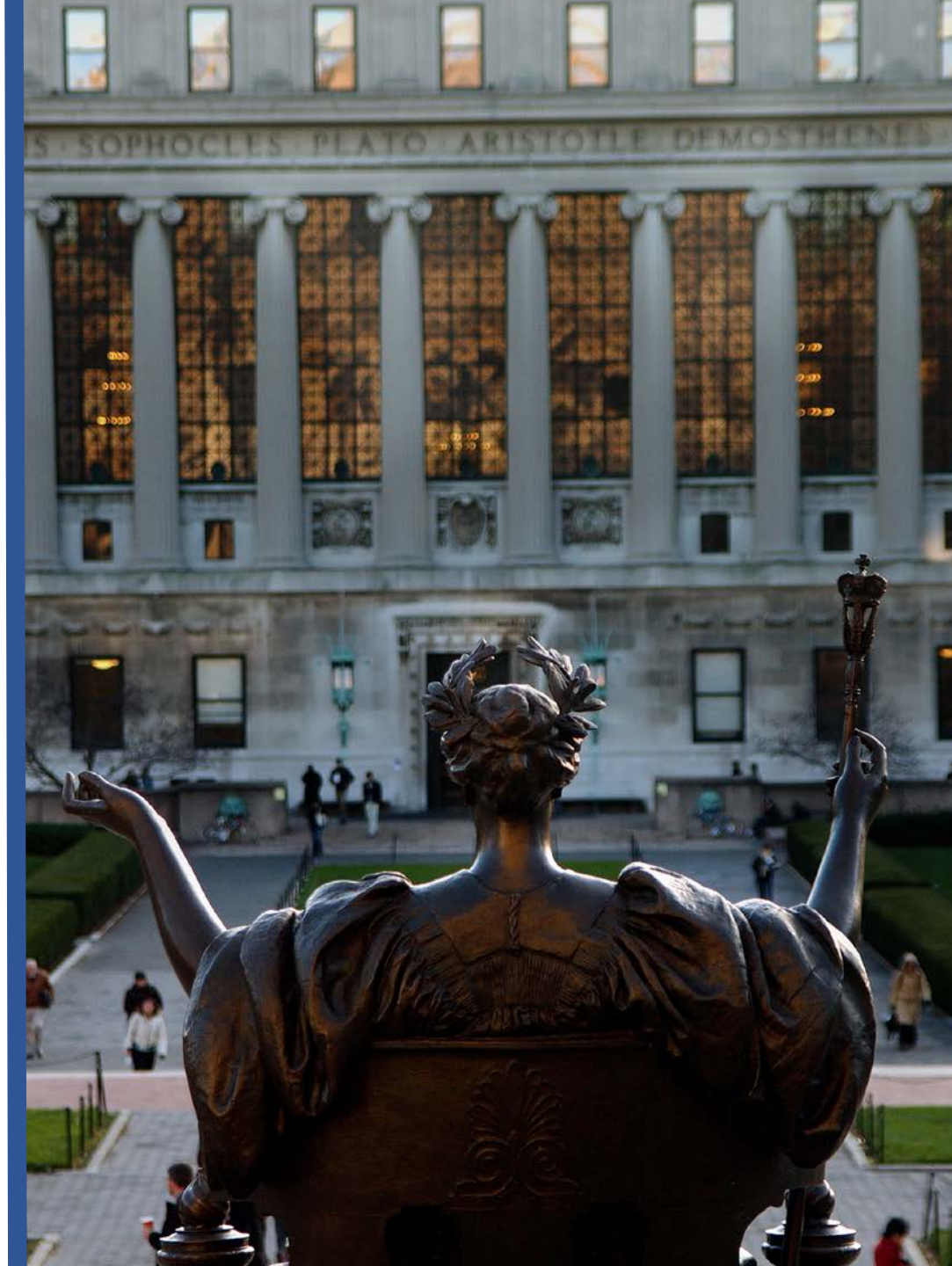
~~Welcome & Introductions~~
~~Habanero Update~~
~~New Cluster Update~~

Assessing Post-Purchase Demand

Chris Marianetti, Chair of SRCPAC

Update from the Training Subcommittee
CUIT Updates
Publications Reporting

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



Assessing Post-Purchase Demand

- **Problems**:
  - Buy-in occurs annually (April-May-June)
  - New Recruits and new requests emerge *anytime*

- **Guiding Questions:**
  - How do we satisfy demand across the academic year?
  - Thoughts?

- **Request:**
  - Communicate with SRCPAC/CUIT early and often
  - Provide us with examples

Spring 2018 Agenda

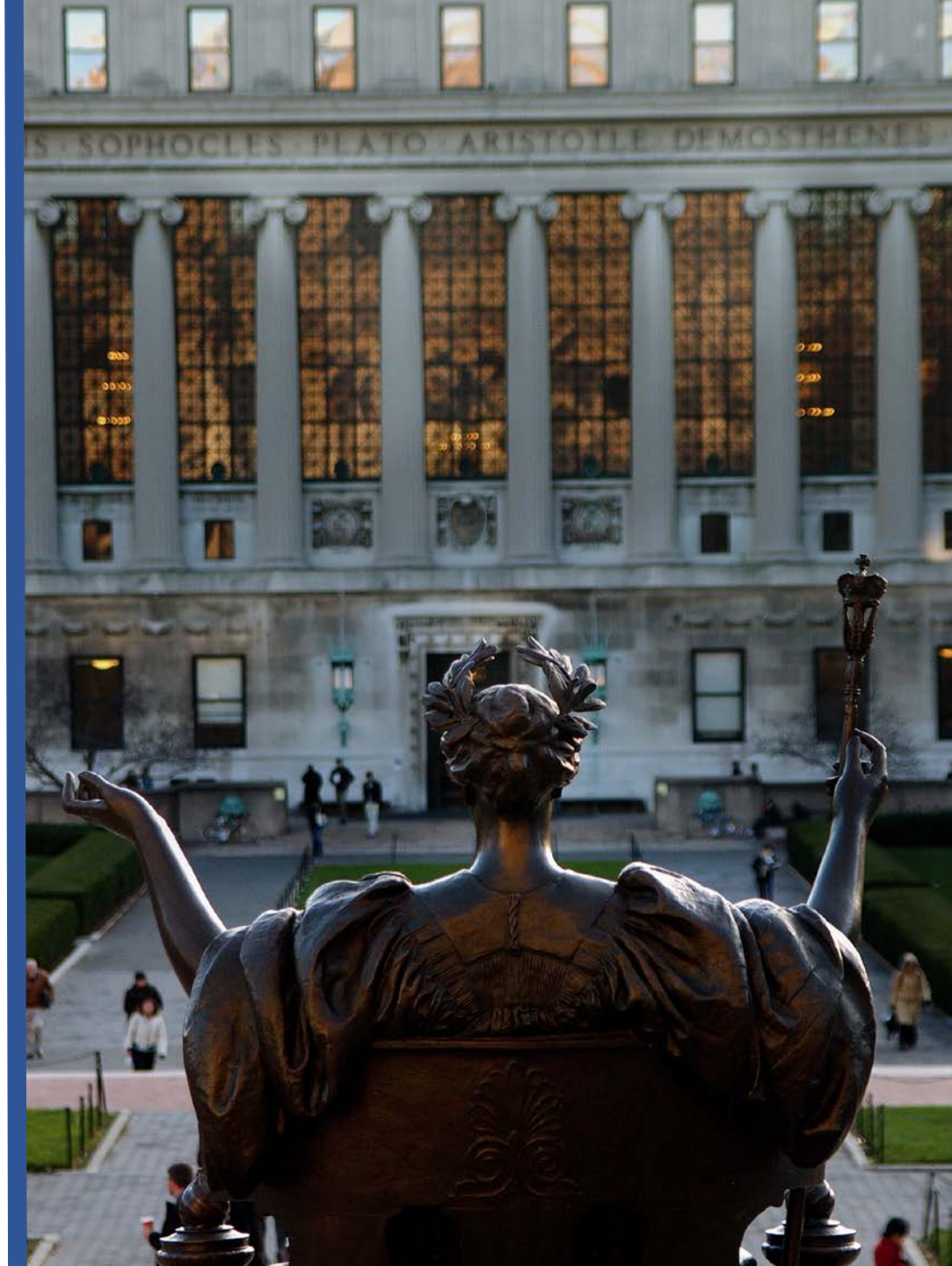
~~Welcome & Introductions~~
~~Habanero Update~~
~~New Cluster Update~~
~~Assessing Post-Purchase Demand~~

Update from the Training Subcommittee

Marc Spiegelman, Chair of Training Subcommittee

CUIT Updates
Publications Reporting

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



15 Subcommittee Members

Marc Spiegelman (Chair)
Departments of Earth and Environmental Sciences and APAM

Ryan Abernathey
Department of Earth and Environmental Sciences

Maneesha Aggarwal
CU Information Technology

Rob Cartolano
Libraries

Halayn Hescock
CU Information Technology

Rob Lane
Department of Computer Science

Kyle Mandli
Department of Applied Physics and Applied Mathematics

Andreas Mueller
Data Science Institute

Barbara Rockenbach
Libraries

Haim Waisman
Department of Civil Engineering and Engineering Mechanics

Christopher Wright (Student Representative)
Department of Applied Physics and Applied Mathematics

Tian Zheng
Department of Statistics/Data Science Institute

Chris Marianetti (Ex Officio)
Department of Applied Physics and Applied Mathematics

Victoria Hamilton (Staff)
Office of the Executive Vice President for Research

Marley Bauce (Staff)
Office of the Executive Vice President for Research

Our Mission

1. Identify current *informal* training in data science and computation
2. Measure demand for new *informal* programs
3. Develop *informal* pilot programs for graduate students
4. Solicit operating budget from internal and external sources
5. *Informal!*

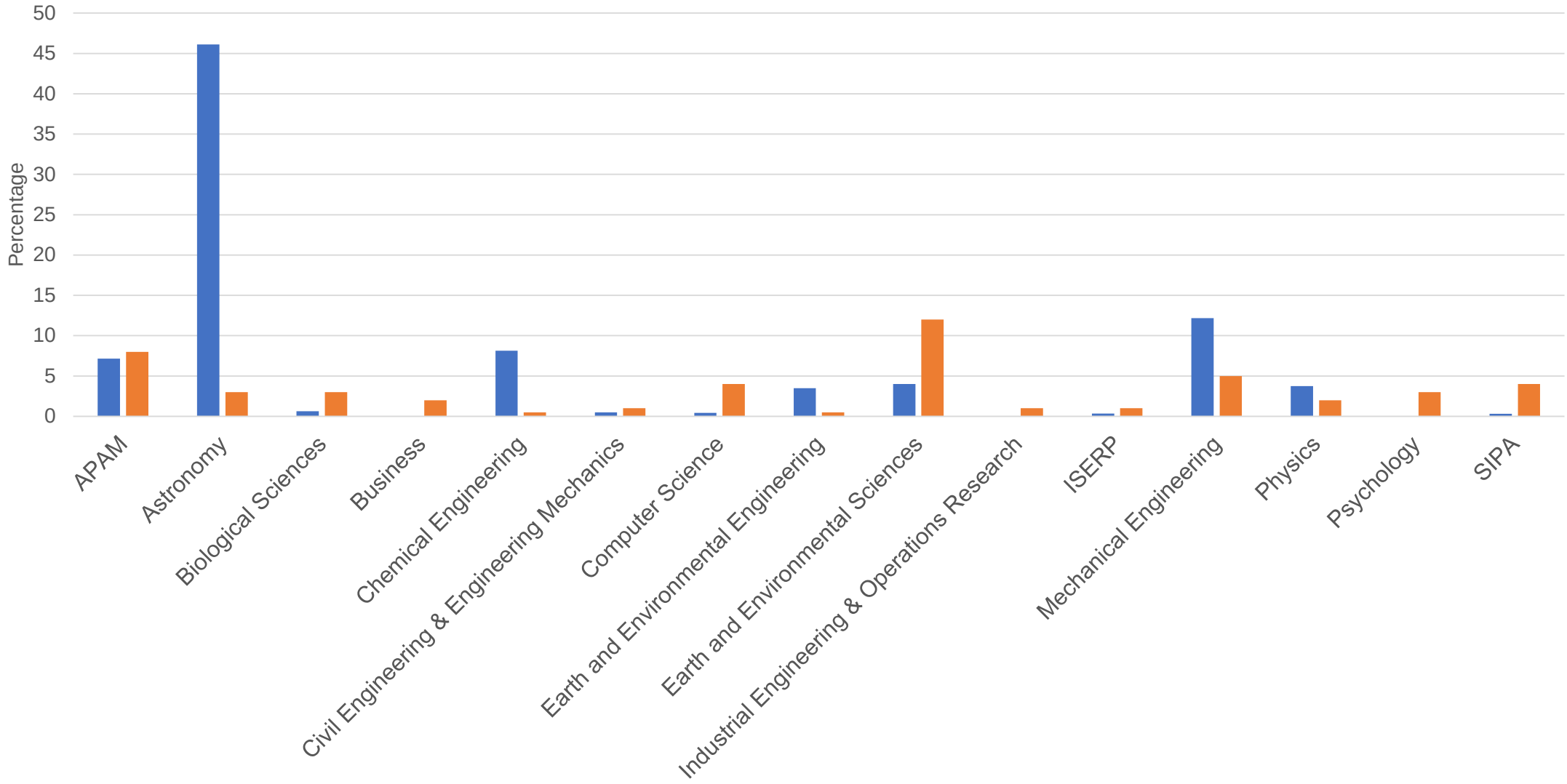
Activity Schedule

Date	Activity
November 2017	Pre-planning meeting
February 2018	Meeting #1
March 2018	Meeting #2
April 2018	Survey to 2,700 graduate students
April 2018	Survey to 12 departments
April 2018	Presentation to SRCPAC
April 2018	Meeting #3
May 2018	Meeting #4
May 2018	Presentation to RCEC

Survey Participation (as of April 13)

- **208 Morningside graduate students**
  - 24 Earth and Environmental Sciences
  - 18 Biomedical Engineering
  - 16 APAM
  - 150 from other Morningside departments
- **6 Morningside departments**
  - 6 in-process

Habanero Usage vs. Survey Participation



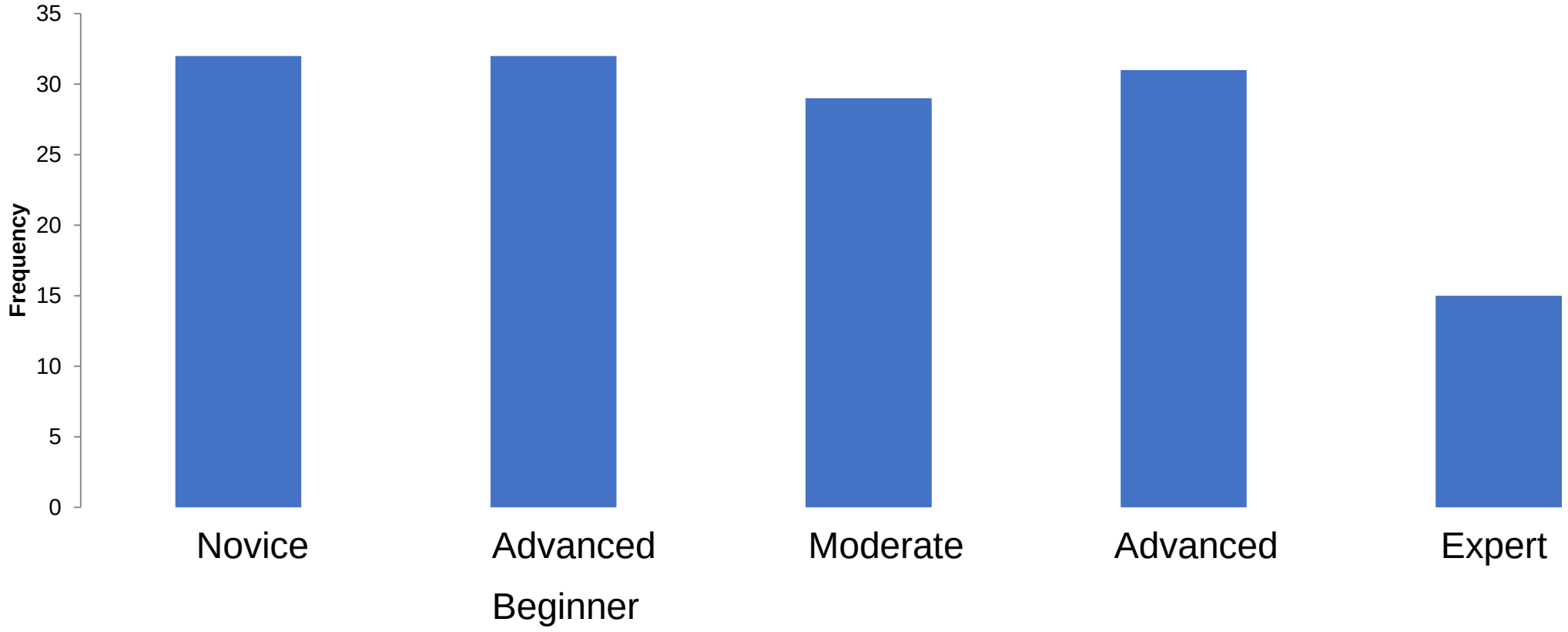
Allow Us to Cherry-pick

Some Initial Findings that Excite Us

Programming Languages Sought

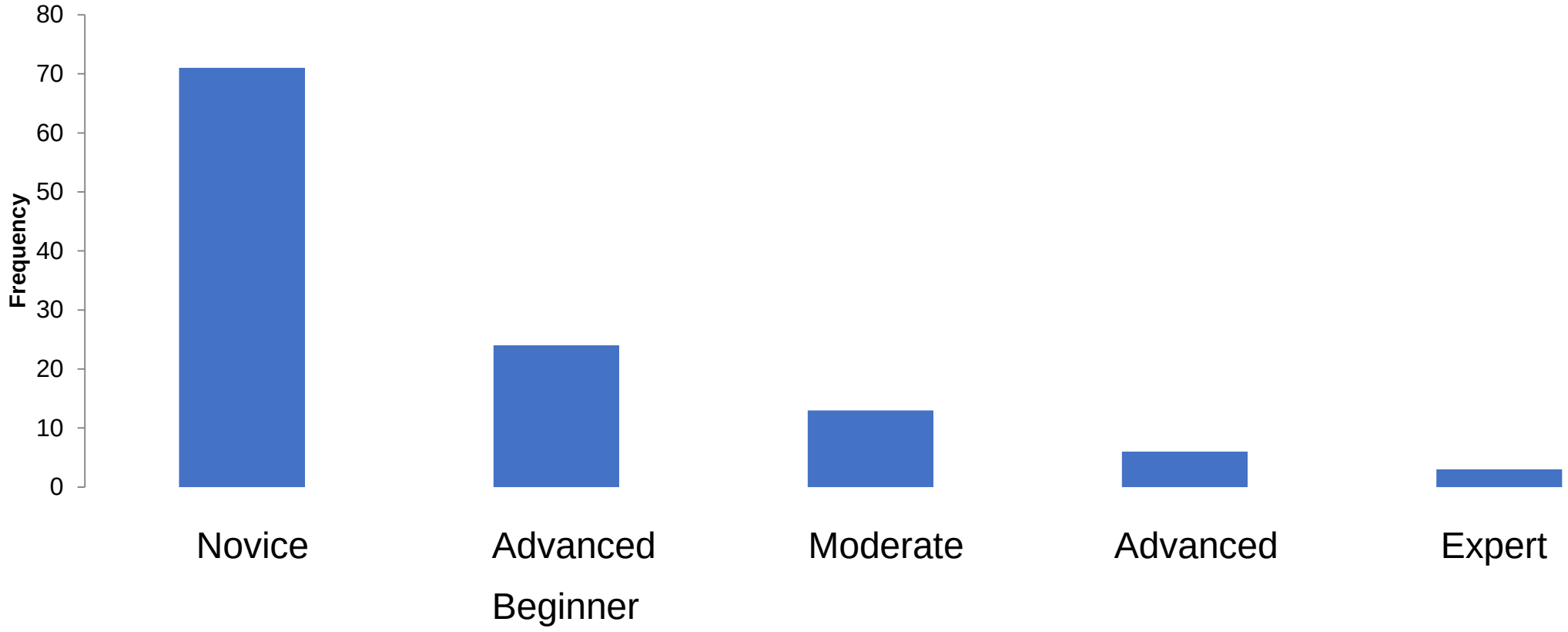
- **Python (38 Times)**
- Julia (8 Times)
- Fortran (7 Times)
- R (6 Times)
- Java (6 Times)
- Matlab (5 Times)

Interesting Tidbits: Python



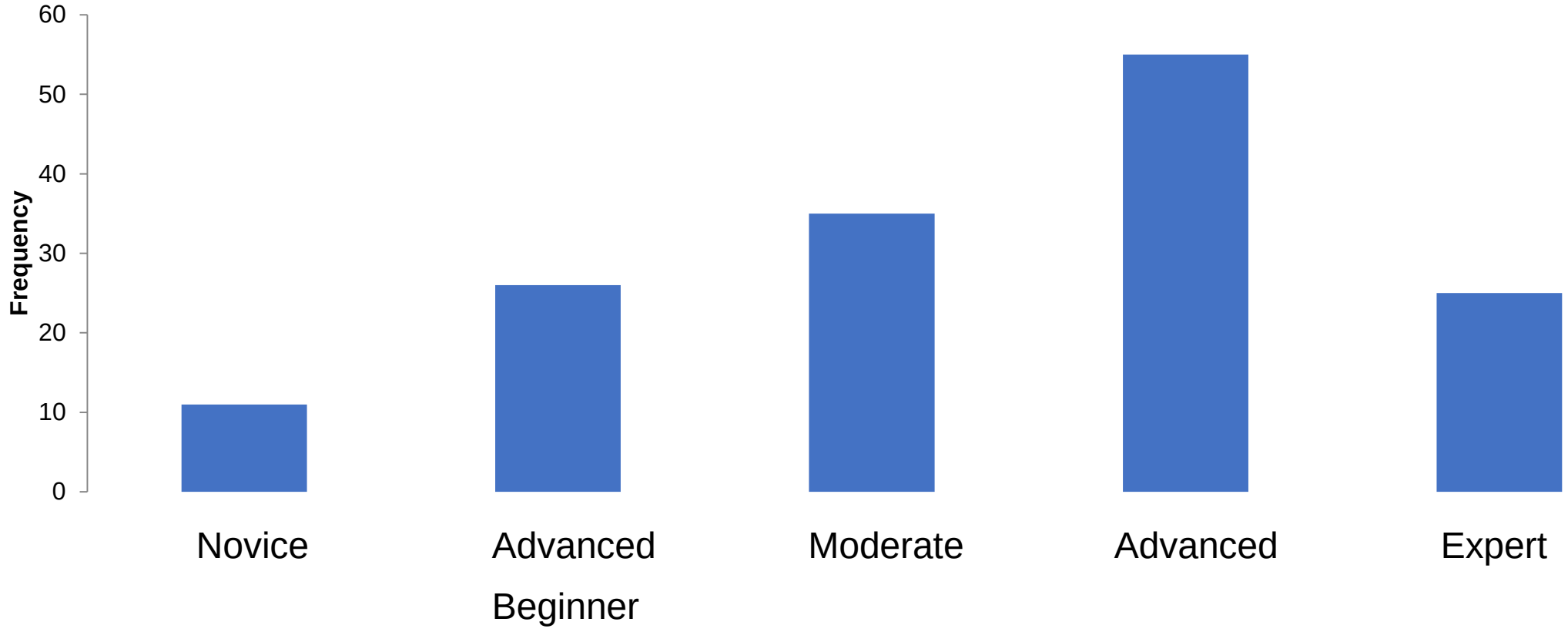
7 Entries for “No Clue”

Interesting Tidbits: Cloud



14 Entries for “No Clue”

Interesting Tidbits: Excel



0 Entries for “No Clue”

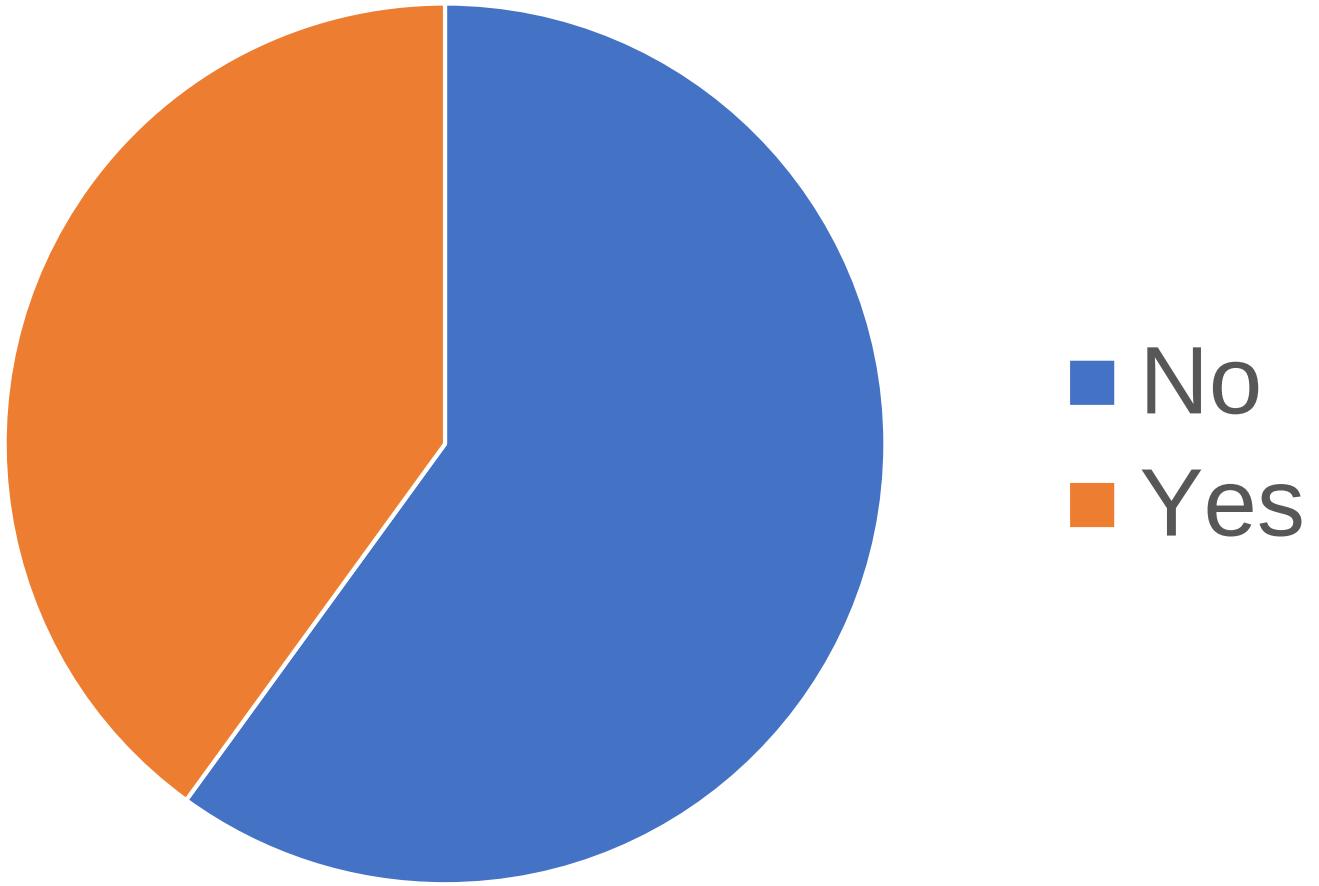
Preferred Informal Trainings

Training Type	Average Score	Number of "Not Sure"
Pre-Semester Boot Camps	3.6	2
Regular Instructional Meetings	3.2	6
Online Self-Study	3.3	2
Mediated Self-Study	3	0
Other	2.8	80
Average	**3.2**	

What Was Hard to Learn?

- **Python (15 Times)**
- Parallel Computing (7 Times)
- Machine Learning (5 Times)
- R (5 Times)
- Cloud Computing (4 Times)
- GIS (3 Times)

Do You Know of Resources?



Universities with Informal Training

- Massachusetts Institute of Technology (13 Times)
- University of California at Berkeley (11 Times)
- Johns Hopkins University (7 Times)
- University of Chicago (4 Times)
- Harvard University (2 Times)
- New York University (2 Times)
- “No” (64 Times)

From the Horse's Mouth

- *"**2-3 hour crash courses** for beginners (Julia, Excel macros, TensorFlow in Python, machine learning packages in R – I would take all of these!)"*

- *"**Encourage training in industry**, bringing back expertise and practices from tech to academia. These days expertise (and funding) lie in the tech giants."*

From the Horse's Mouth

- *"I just want to emphasize that **online tutorials and mediated self-study would be entirely unhelpful**."*

- *"**Short workshops or classes on specific topics during the semester** would also be very useful. Like a 1 week course on the basics of how to use SQL, or a 2 week course on Julia aimed at students who already have some prior knowledge in another programming language."*

Additional Thoughts?

- *"While my job does not enable me to attend weeklong boot camps or commit to a weekly course, I can and would avail myself of opportunities to attend a* ***few hours of crash-course style seminar instruction*** *– just enough to get comfortable continuing to teach myself – in new data science languages and techniques every few weeks."*

Additional Thoughts?

- *"**Sustained support** is what most students need... as a graduate instructor who has taught R and STATA, the short-spurts of lab time over one semester are not enough at all for students to gain proficiency. What would really help students is to have **staff in the libraries** with deep knowledge of the applications and types of problems they need to solve, such that they can slowly build their knowledge over their years at Columbia across classes and projects."*

Brainstorm for New Programs

- Weeklong boot camp for incoming graduate students
  - Recorded, curated, and published by Columbia Video Network (CVN)
- Software Carpentry institutional partnership and boot camps
- Distinguished lectures by industry programmers
- Curated online training modules
- Drop-in room for research computing
- Full-time coordinator (reporting line TBD)

Help Us Brainstorm?

Tell Us Your Ideas for *Informal* Training Programs

COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

How We Ask for Money

- **Internal**: Budget request from RCEC (May presentation)

- **External**: NSF Partnerships between Science and Engineering Fields and the NSF TRIPODS Institutes (TRIPODS+X)
  - TRIPODS: Transdisciplinary Research in Principles of Data Science
  - Theoretical data science applied to non-traditional discipline
  - Focus on informal curriculum development
  - $200k for 3 years
  - May 29 deadline

Contact Us *Anytime*

Marc Spiegelman, mspieg@ldeo.columbia.edu
Marley Bauce, mb3952@columbia.edu

COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

Spring 2018 Agenda

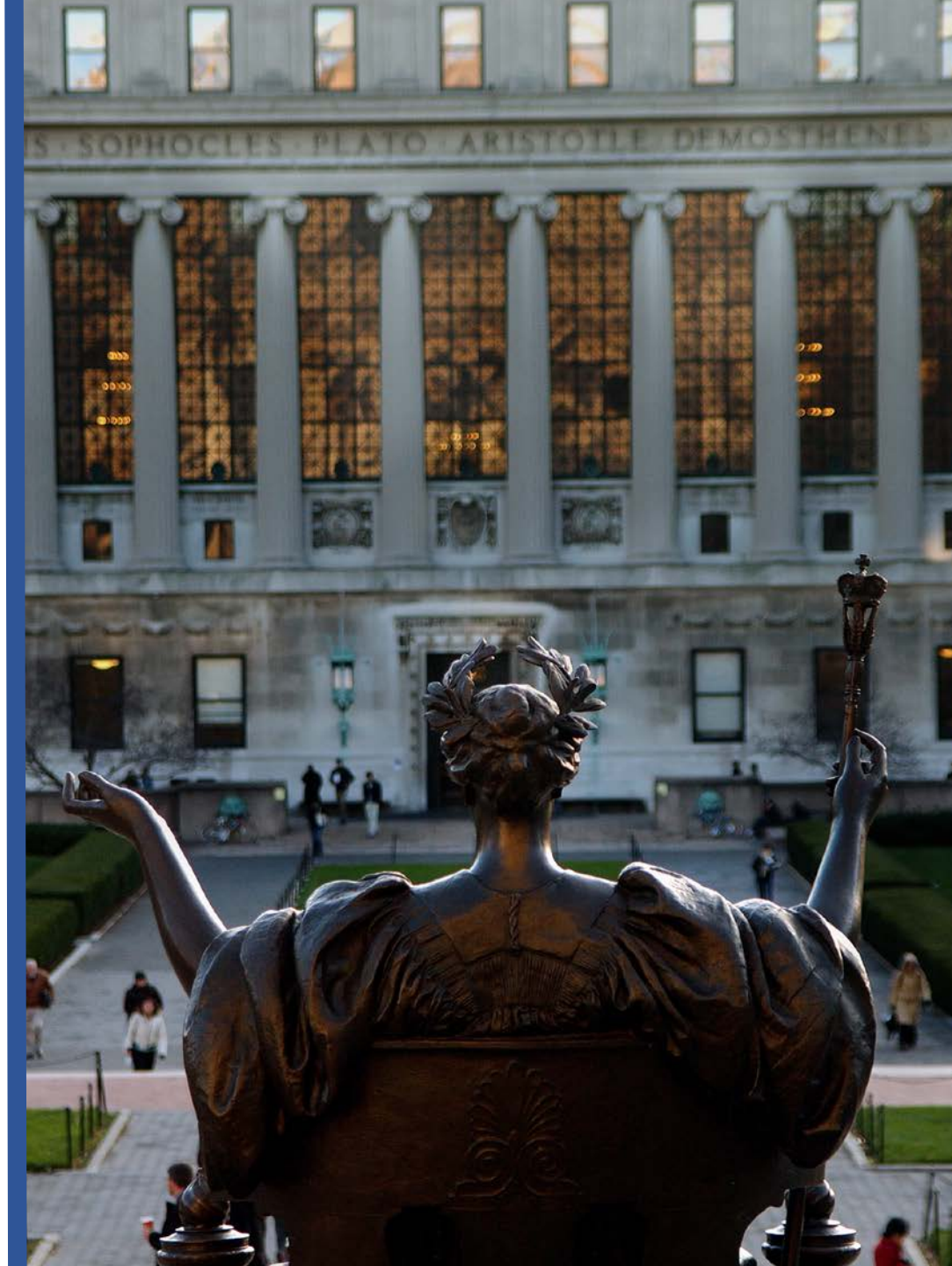
~~Welcome & Introductions~~
~~Habanero Update~~
~~New Cluster Update~~
~~Assessing Post-Purchase Demand~~
~~Update from the Training Subcommittee~~

CUIT Updates

George Garrett, Manager of Research Computing Services
Sander Antoniades, Lead Research Systems Engineer
Maneesha Aggarwal, AVP for Academic IT Solutions
Peter Jorgensen, Lead Infrastructure Systems Engineer

Publications Reporting

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



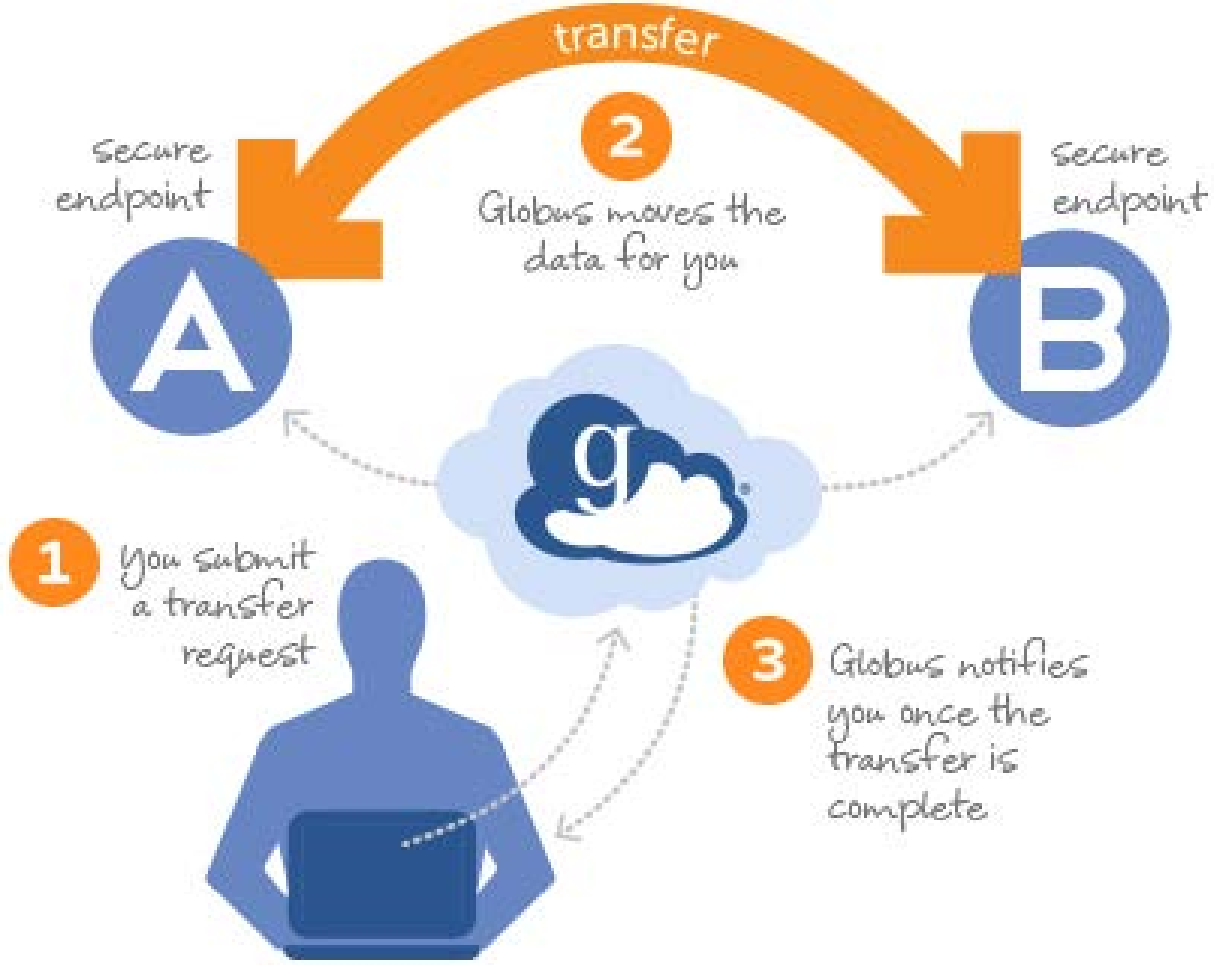
Secure Data Enclave

- Secure remote desktop to store and analyze restricted data:
  - Personally Identifiable Information (PII)
  - Protected Health Information (PHI)
  - Payment Card Information (PCI)

- Launching Spring 2018

Globus

- Provides secure, unified interface to research data.
- "***Fire and Forget***" high-performance data transfers between systems within and across organizations.
- Procuring an enterprise license.



Emerging Technologies

Maneesha Aggarwal
Assistant Vice President for Academic IT Solutions
CU Information Technology

COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

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)
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

Business Associate Agreement

- The University's BAA with Amazon allows PHI to be stored and/or used with resources running in AWS
- BAA requires opting-in the specific account
- Some caveats: only specific AWS services are covered
- All other applicable Columbia and CUMC policies remain in effect, such as RSAM registration
- https://aws.amazon.com/compliance/hipaa-eligible-services-reference/

AWS Direct Connect

- Activated January 2018, currently piloting
- 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
- **Pricing not finalized, likely $200–250/month** (including VPN failover)

Why Direct Connect?

- **Egress costs reduced** from 9¢/GB to 2¢/GB (⬇ 75%)
- More consistent performance
- Directly communicate with resources using private IPs (i.e., 10.193.255.100)
- Traffic never touches public Internet

More Information

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

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

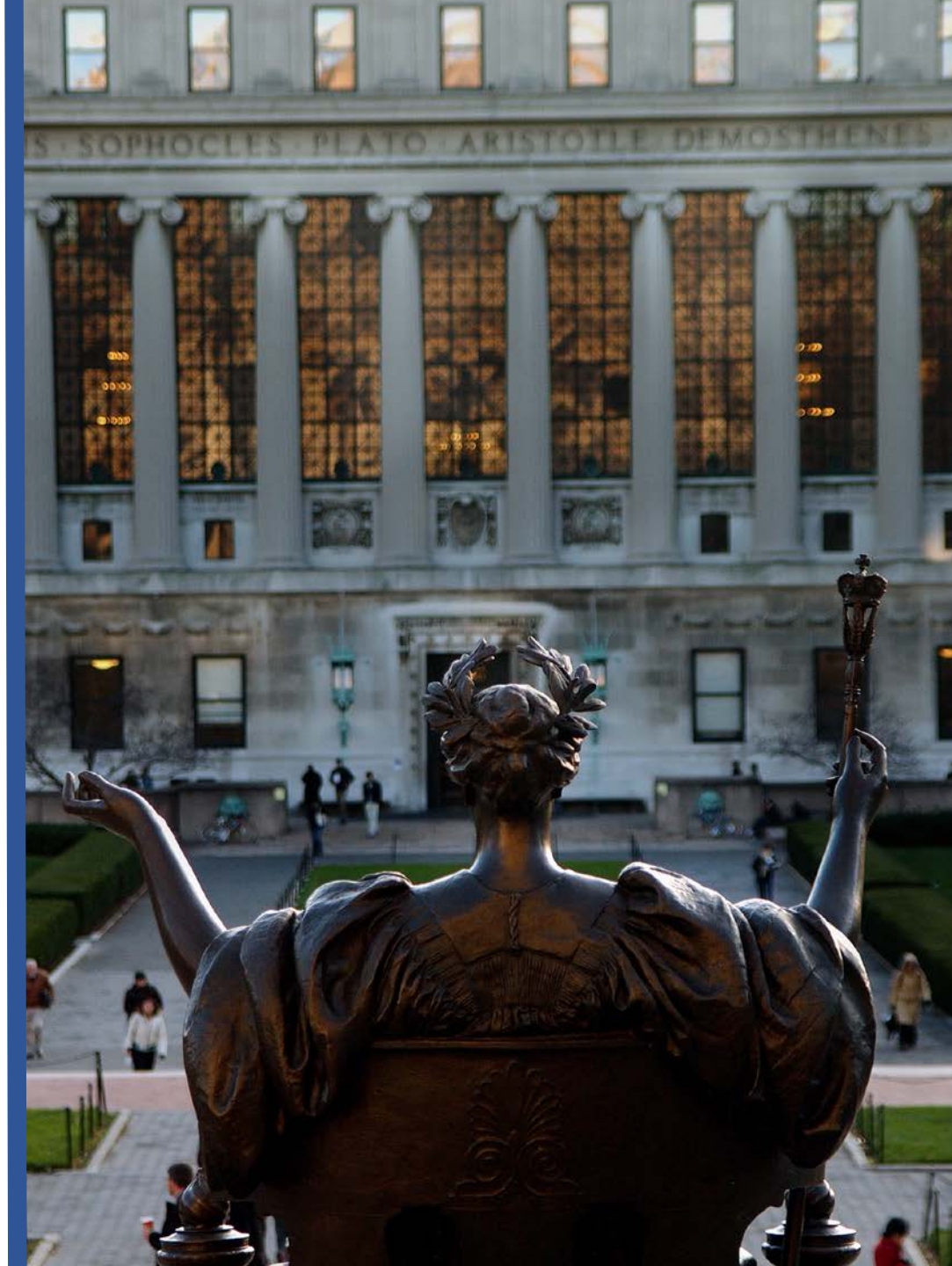
Spring 2018 Agenda

~~Welcome & Introductions~~
~~Habanero Update~~
~~New Cluster Update~~
~~Assessing Post-Purchase Demand~~
~~Update from the Training Subcommittee~~
~~CUIT Updates~~

Publications Reporting

Chris Marianetti, Chair of SRCPAC

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee



Reporting Publications

- Research Computing Executive Committee in **five weeks**
- Expects report of *all* Yeti- and Habanero-supported publications
- Publication quantification strengthens case for continued support
- Email from **srcpac@columbia.edu** requesting updates – please respond!
- Libraries investigating DOI add-in for all publications

Thank You!

See You in the Fall

COLUMBIA | RESEARCH
Shared Research Computing
Policy Advisory Committee