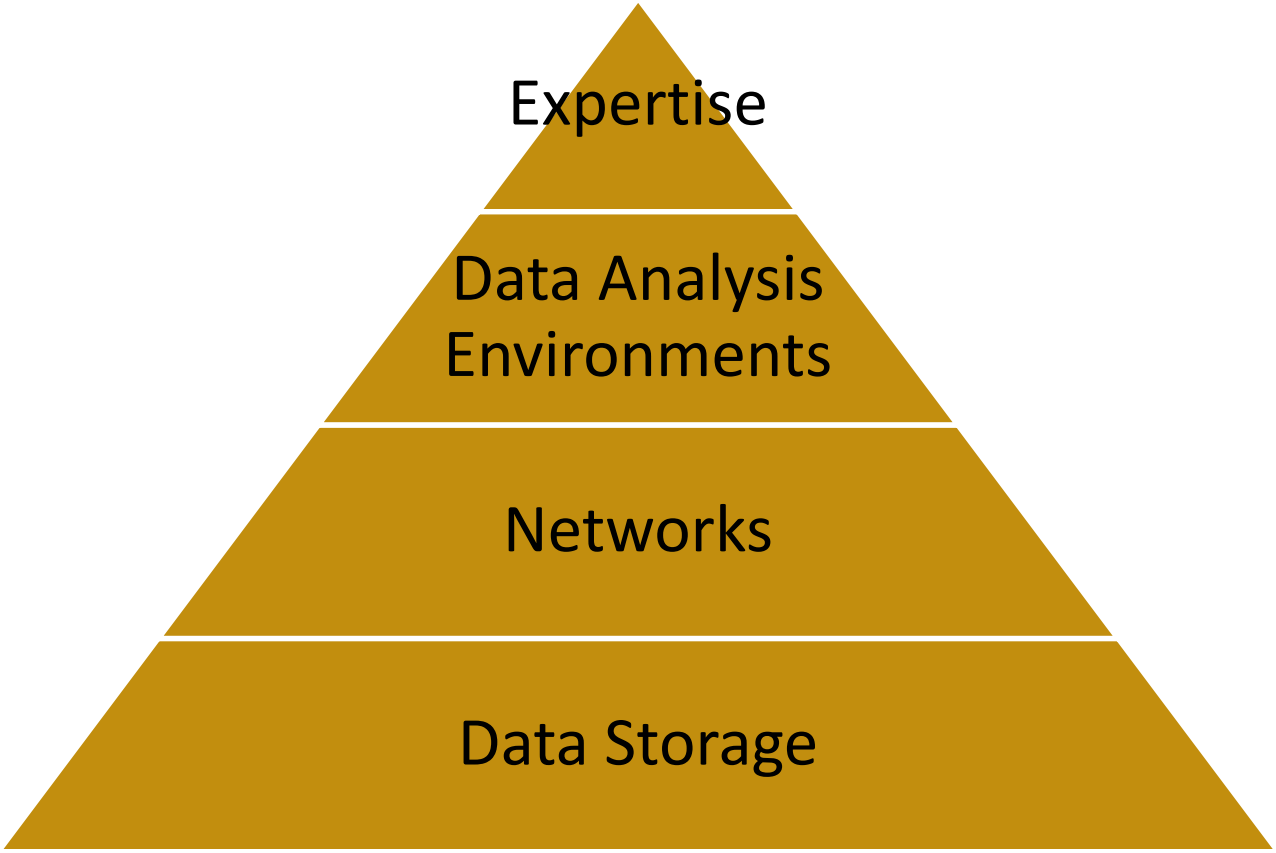
COMPUTING RESOURCES FOR DATA SCIENCE

PRESTON SMITH

DIRECTOR, RESEARCH SERVICES AND SUPPORT

PURDUE UNIVERSITY

Hierarchy of Data Science Needs



PURDUE UNIVERSITY

Foundation of Data Science

Data

- One cannot do data science without data!

PURDUE UNIVERSITY

Common Needs

Almost every scientist has the same general data needs

- Take or store research data
- Move research data
- Collaborate within the lab
- Simulate and analyze data
- Share research data outside the lab

PURDUE UNIVERSITY

Data Storage is Multi-dimensional

Multi-dimensional Intersection of:

- Data Size
- Usage pattern
- Accessibility requirements
- Data/Compute locality
- Sharability
- Etc

Example: < 10MB Office documents, with collaborative editing, shareability off-campus, no need for access from HPC

Vs: 10TB dataset generated from instrument, processed on HPC, private to research lab

PURDUE UNIVERSITY

WORKING STORAGE

PURDUE UNIVERSITY

On-Campus Research Storage Solutions

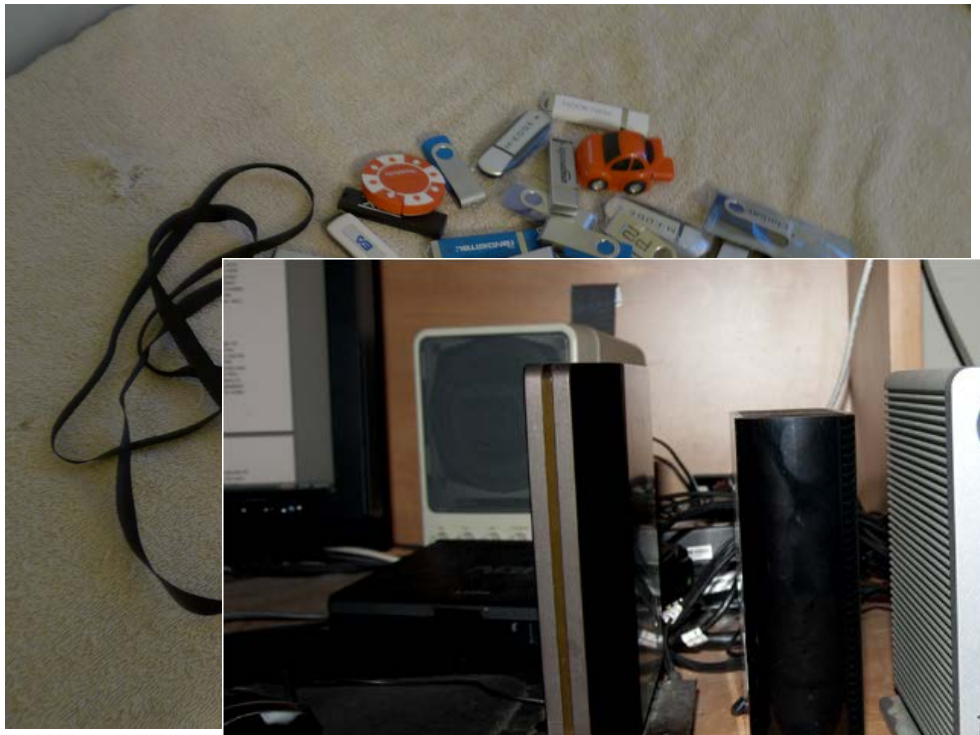
Some are fairly reasonable

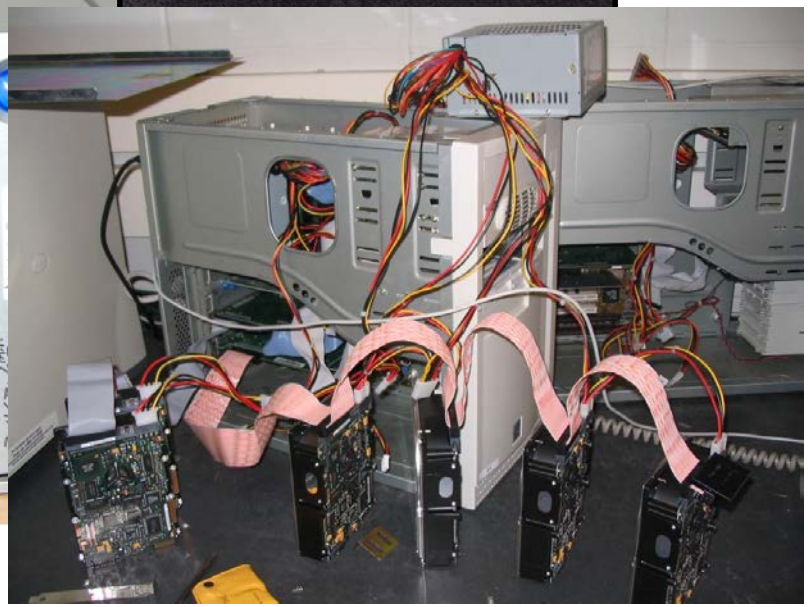
- Left on your own, you will find a way!

PURDUE UNIVERSITY

On-Campus Research Storage Solutions

Others Less So





On-Campus Research Storage Solutions

Archival Storage

PURDUE UNIVERSITY

Storage Gaps on Community Clusters

Over 20% of 2014 research tickets involved data

Many Common Requests

- I need more space than I can get in scratch
- Where can I install applications for my entire research lab?
- I'm ***actively working*** on that data/software in scratch:
  - I have to go to great lengths to keep it from being purged.
  - I shouldn't have to pull it from Fortress over and over
- Can I get a UNIX group created for my students and I?
- Is there storage that I can get to on ***all*** the clusters I use?
- I have funding to spend on storage – what do you have to sell?
- I need storage for my instrument to write data into
- My student has the only copy of my research data in his home directory, and he graduated/went off the grid!

PURDUE UNIVERSITY

The Research Data Depot

At $70 per TB/year

- Storage oriented around your research lab, with
  - Snapshots
  - Multi-site copies of your data
  - Disaster protection
  - A scalable, expandable storage resource optimized for HPC
- Access to Globus data transfer service, and endpoint sharing

PURDUE UNIVERSITY

The Research Data Depot

Impact

- Over 562 research labs are Depot partners!
  - ***60% are not HPC users!***
  - *Thousands of individual users*
- Almost all storage (>2 PB) sold!
- A research group purchasing space has purchased, on average, nearly 10 TB.
- Other institutions looking to Purdue for leadership from our Depot storage service

PURDUE UNIVERSITY

The Research Data Depot

The Service

- It's important to think of Depot as a "data service" – not "storage"
- It is not enough to just provide infrastructure
  - "Here's a drive, have fun"

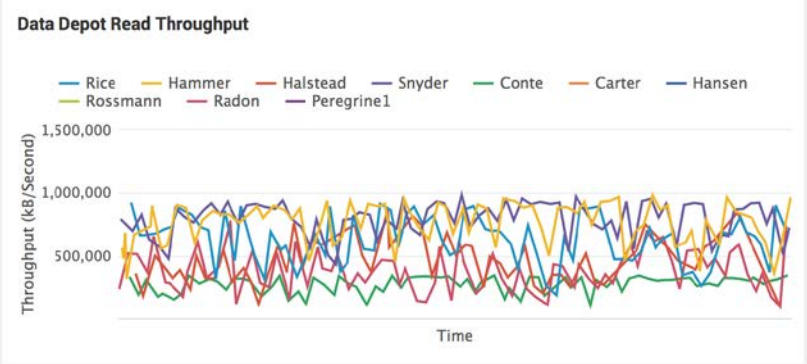
- Our goal: enabling the frictionless use and movement of data
  - *Instrument -> Depot -> Scratch -> Fortress -> Collaborators -> and back*
  - Continue to improve access to non-UNIX users

PURDUE UNIVERSITY

Depot Strengths and Weaknesses

Strengths

- Scalable
- Very local to computing resources
- It's a regular POSIX filesystem
- Aimed at the individual PI
- Built around common data management patterns of research labs

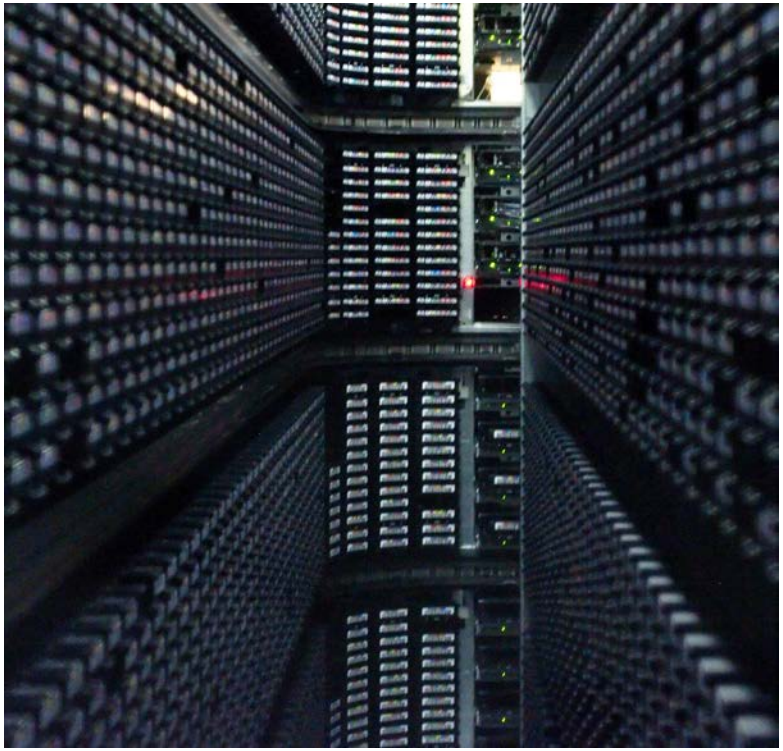


PURDUE UNIVERSITY

ARCHIVAL STORAGE

PURDUE UNIVERSITY

Fortress – Archival Storage



Write once, read never

- Long-term archival storage for research data
- Value-add offering to HPC or research storage users

PURDUE UNIVERSITY

Fortress Archive

Characteristics

- No Quota
- Data protection by multiple tapes
- Tapes allow for transparent, continually-modernized media
- More latency than spinning disk
- Ideal for original copies of data, final products.
  - Infrequently accessed!
- Strongly prefers large files vs small ones
  - Zip up big directories!
- Limited access methods - “put” or “get”
  - Globus, HSI, HTAR

PURDUE UNIVERSITY

Fortress Archive

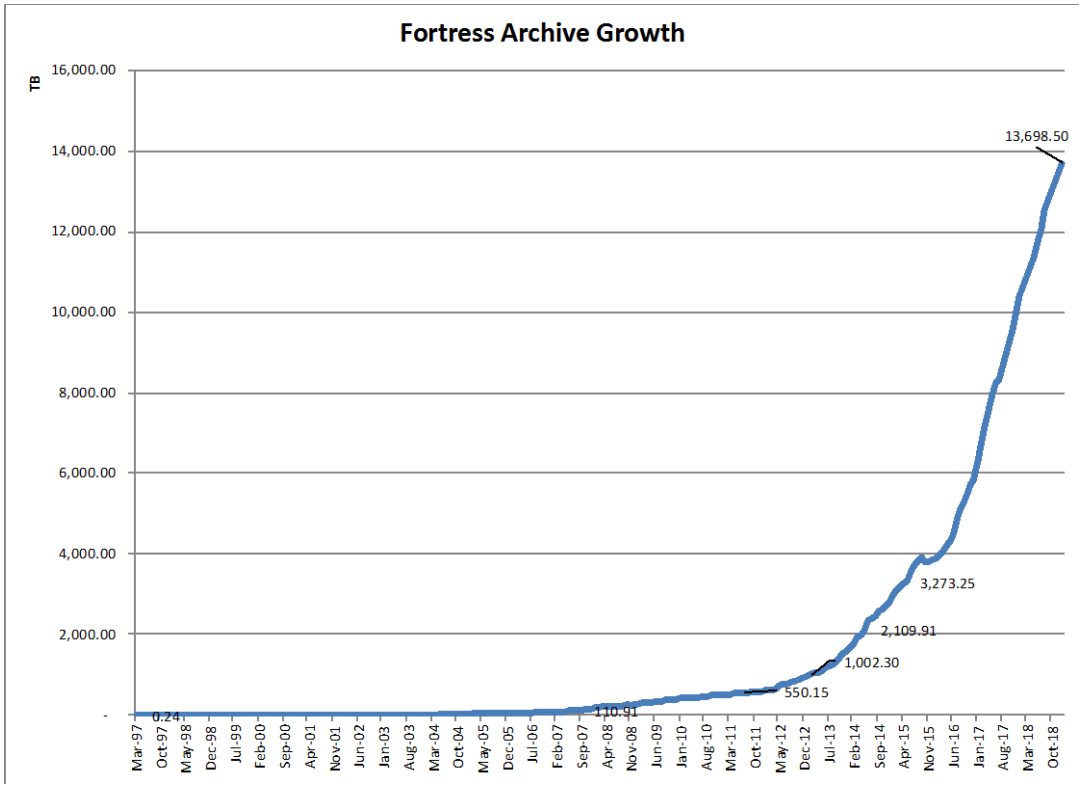
What is Fortress not?

- Fortress is not a substitute for PURR!
- Just data storage, no curation, preservation, publishing.

- Fortress is not a network drive
- Not a backup system
- Not a network share to map from your PC

PURDUE UNIVERSITY

Fortress Archive



Vital Statistics

- ~1200 users
- Median user stores 214 GB

Costs?

- Included with your cluster/Depot purchase
- Chargeback for using more than .5PB

PURDUE UNIVERSITY

Sharing and Transfering Big Data

Transfer and share large datasets….

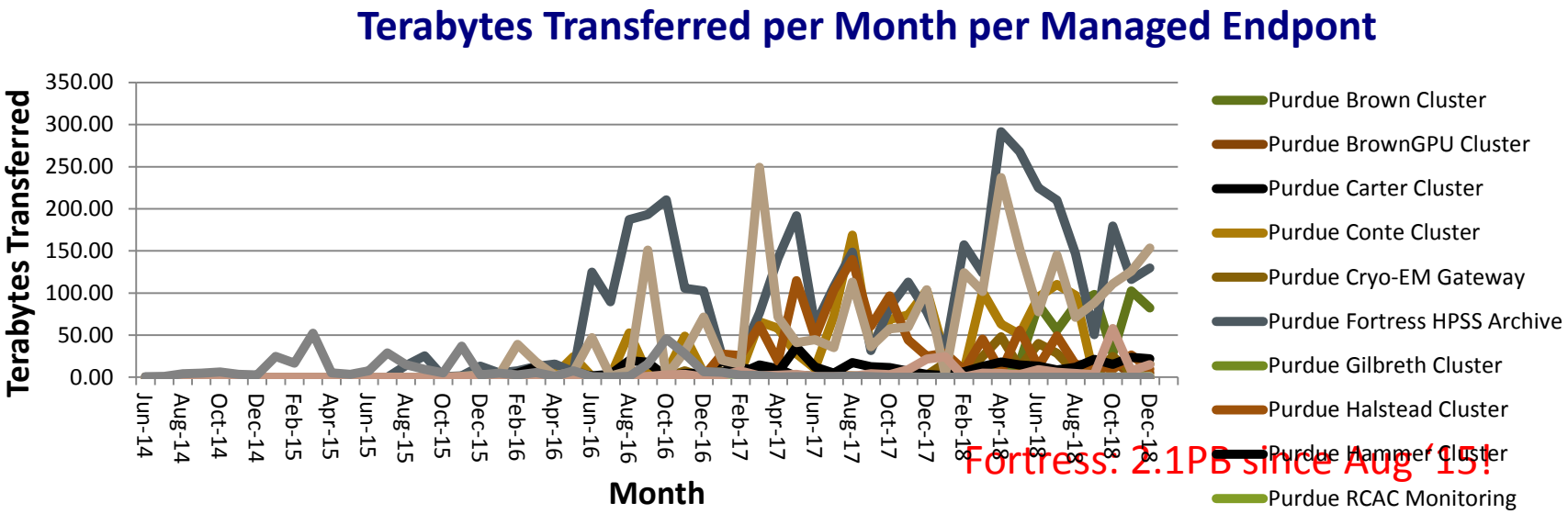
…. With dropbox-like characteristics ….

https://transfer.rcac.purdue.edu

…. ***Directly from your own storage system!***

PURDUE UNIVERSITY

Globus



Fortress: 2.1PB since Aug '15!

- 2018:
- 3.1 PB transferred (> 400TB in March!)
- Average of 266 TB, 123 unique users per month

PURDUE UNIVERSITY

REGULATED DATA

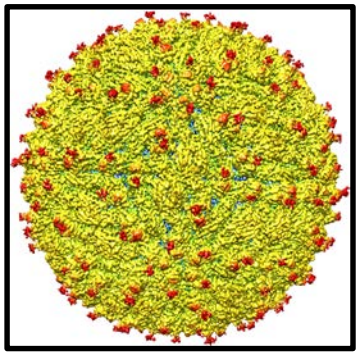
PURDUE UNIVERSITY

Regulated Data

Not everybody has open data

- We can support export-controlled research fairly well today
  - **ITAR, CUI in "REED-GovCloud"**
- And are actually well-positioned as a national leader in this space, and capabilities continue to improve.

- HIPAA data, however, is historically been a challenge
- Regulation-compliant collaboration and sharing is a challenge



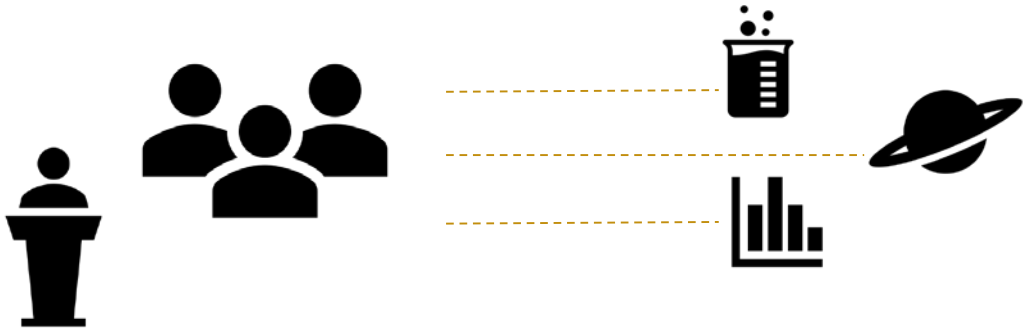
PURDUE UNIVERSITY

REED+

REED+

A managed research ecosystem with storage, high speed computing capability and security to efficiently and cost effectively handle Purdue's controlled research data and processing needs with the same scalability and cost-effectiveness of open data services.

2 year, $600k NSF Award
#1840043 to build framework for research cybersecurity



PURDUE UNIVERSITY

Box

Available now!

- Unlimited Cloud Storage for your research group via your Purdue account.
- Easy sharing with collaborators
- Access from anywhere, anytime
- Collaborate in real-time

- ”REED Folder” available for HIPAA or FERPA research projects today.
- Future regulations under investigation

- **Talk to us about cloud storage for your research group!**

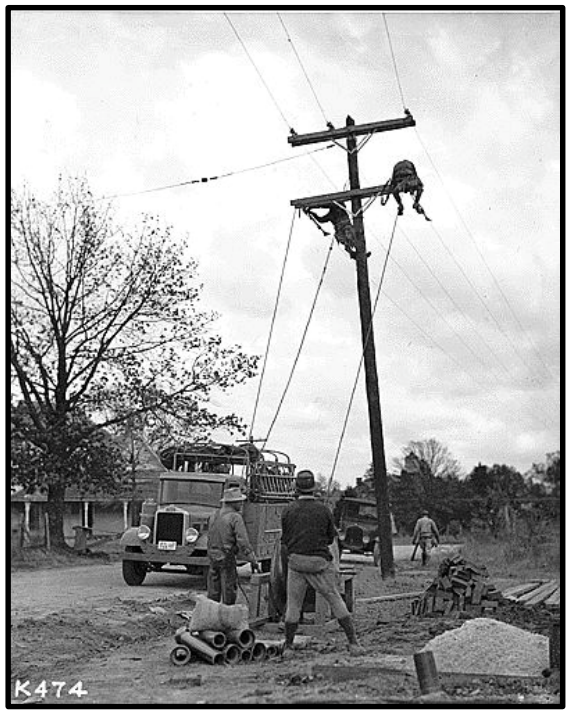
PURDUE UNIVERSITY

DATA NETWORKING

PURDUE UNIVERSITY

Networking

The Last Mile



- The last mile is a key piece in the connectivity puzzle!
- Fast, friction-free networks necessary to get data from big-data hot spots!

2 year, $325k grant from NSF to enhance connectivity to key instruments

PURDUE UNIVERSITY

DATA ANALYSIS

PURDUE UNIVERSITY

Computing for the Long-tail

Data Analysis Tools

- For many, your laptop or PC may be all you need to meet your data science needs.
- Add Data Depot, Fortress, or Box Research Folders, and you may be all set!

- **Until**, you're suddenly not.

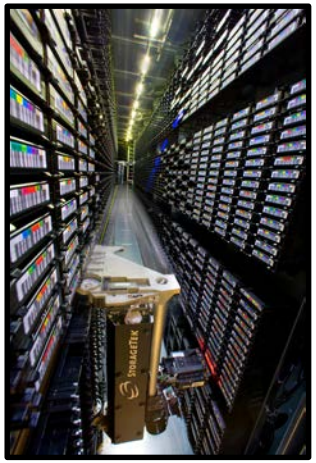
Does your laptop shape the research questions you're asking?

PURDUE UNIVERSITY

Data Workbench

When your laptop is no longer sufficient

Get shared, interactive computing on supercomputer-grade hardware



MATLAB

RStudio

jupyterhub

ArcGIS

MySQL

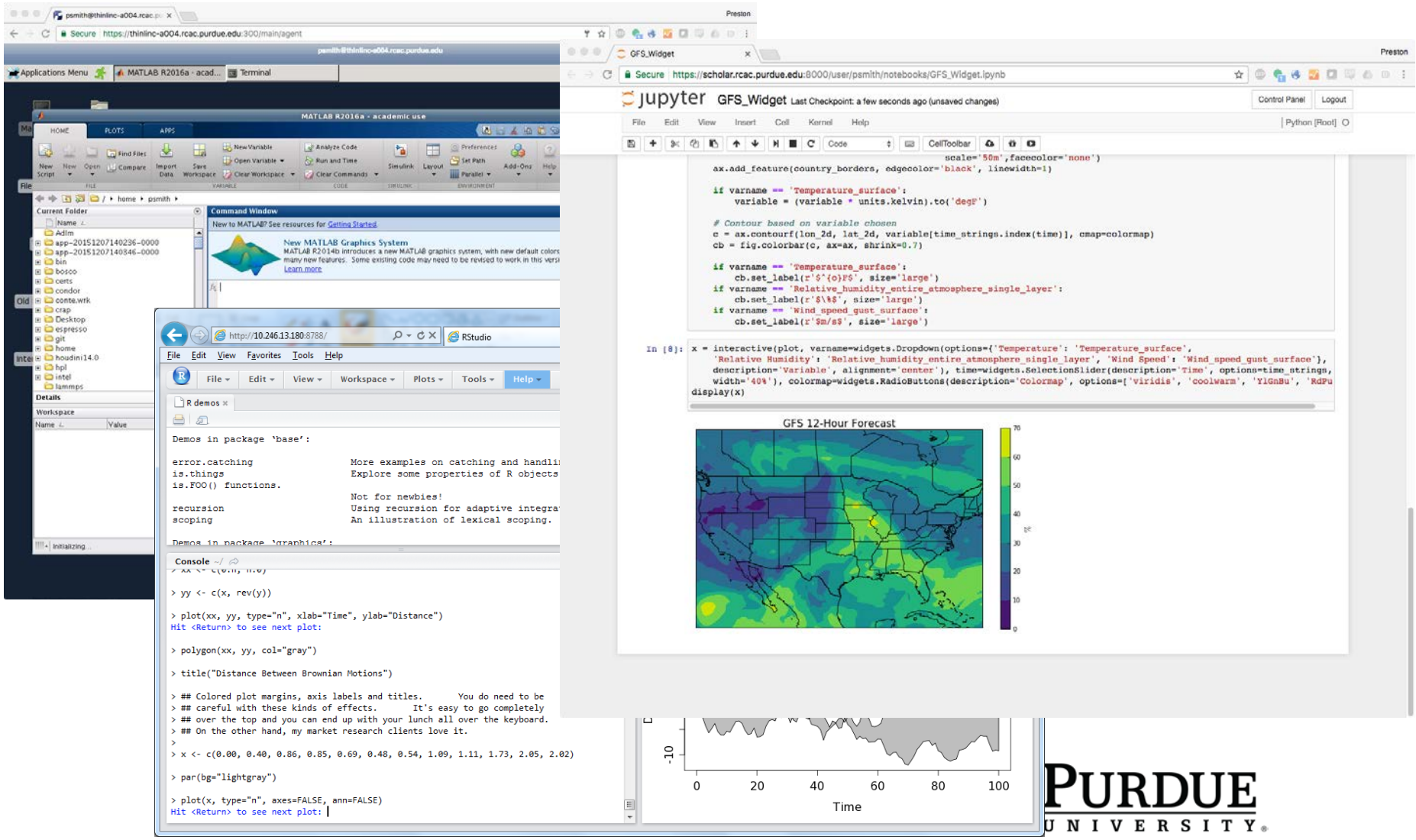
globus

APACHE Spark

git

PURDUE UNIVERSITY

Interactive Computing from Your Browser



PURDUE UNIVERSITY

Data Workbench

Service

- $300 annual charge for access
- Easy access to data analysis tools
- Run virtual private Windows desktops
  e.g. for GIS software
- Run virtual private Linux images in containers
- Integrate with RCAC services
  - Depot, Fortress, Globus, Github, Self Service, etc.
- Grow to batch HPC as your needs grow!
- Add same interactive capabilities to community cluster frontends

PURDUE UNIVERSITY

COMMUNITY CLUSTERS

PURDUE UNIVERSITY

Research Computing Background

Linux Clusters

- By the early 2000s, Purdue had built our first Linux clusters, and a general purpose resource made from retired computer lab systems was in production.

But:

PCs are not built to do HPC
Already 3+ years old by the time they become cluster nodes!

... And at the time, privately-run clusters were proliferating in labs and newly-made datacenters across campus!

PURDUE UNIVERSITY

Community Cluster Program

Principles

- Centralizing HPC yields institutional benefits
  - Coalition of the willing
  - Sharing costs in partnership with faculty
  - Economies of Scale
  - Maximize utilization
- We build a cluster each year
  - Cluster is a service, buying in provides 5 years of access.
- Partners get out at least what they put in
  - Buy 1 node or 100, you get a queue that guarantees access up to that many CPUs

PURDUE UNIVERSITY

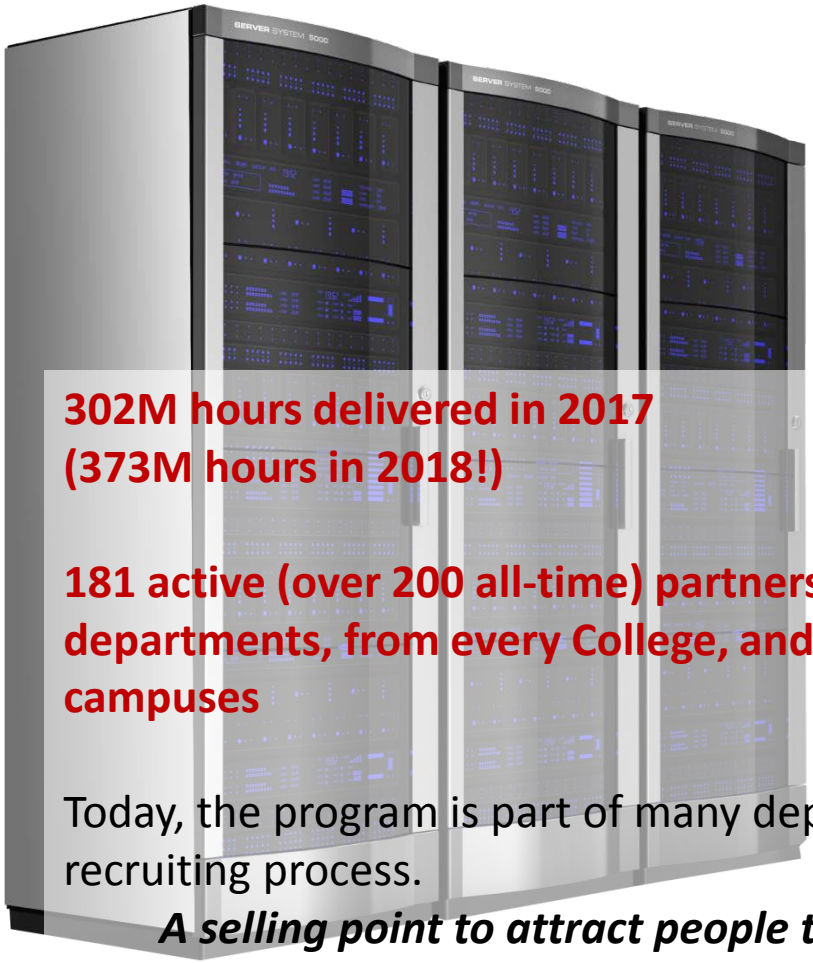
Community Cluster Program

The Rules

- But wait, there's more!!
  - What if your neighbor isn't using their queue?
    - You can use it, but your job has to run in 4-hour chunks if they want to run.
- You don't have to do the work
  - Your grad student gets to earn their PhD rather than run your cluster.
    - Nor do you or your department head have to provide space in your lab for computers.
  - ITaP provides storage, data center space, systems administration, application support.
  - Just submit jobs!

PURDUE UNIVERSITY

Cluster Program Partners



302M hours delivered in 2017
(373M hours in 2018!)

181 active (over 200 all-time) partners from 54 departments, from every College, and 3 Purdue campuses

Today, the program is part of many departments' faculty recruiting process.

A selling point to attract people to Purdue!

$10M invested since 2012

CMS Tier2	14992
Mechanical Engineering	12884
Electrical and Computer Engineering	11668
Aeronautics and Astronautics	10712
Earth, Atmospheric, and Planetary Sciences	6060
Materials Engineering	4704
ITaP	3824
Chemistry	3144
Nuclear Engineering	2044
Chemical Engineering	2008
Biological Sciences	1452
Physics and Astronomy	1404
Medicinal Chemistry and Molecular Pharmacology	1320
Biomedical Engineering	1168
Agricultural and Biological Engineering (Biological Engineering)	1056
Statistics	944
Industrial Engineering	824
Mathematics	760
Other Executive Vice President for Research and Partnerships	696
Civil Engineering	680
Agronomy	620
Computer Science	504
Bioinformatics Core	456
Industrial and Physical Pharmacy	424
Computer and Information Technology	408
Composites Manufacturing and Simulation Center	400
Biochemistry	324
Forestry and Natural Resources	232
Cancer Center	184
Horticulture and Landscape Architecture	168
NULL	120
Animal Sciences	116
Health Sciences	68
Botany and Plant Pathology	64
Entomology	64
Agricultural and Biological Engineering (Agricultural Systems Mgmt)	60
Computer Graphics Technology	56
School of Engineering Technology	48
Brian Lamb School of Communication	40
Agricultural Economics	36
Economics	24
Management	24
Purdue Institute of Inflammation, Immunology and Infectious Disease	24
Food Science	20
Other College of Agriculture	20
Other Libraries	20

10 HPC SYSTEMS

STEELE

7,216 cores, Installed May 2008

Retired Nov. 2013

COATES

8,032 cores, Installed May 2008

24 departments, 61 faculty investors

Retired Sep. 2014

ROSSMANN

11,088 cores, Installed Sept. 2010

17 departments, 37 faculty investors

Retired Sep. 2015

HANSEN

9,120 cores, Installed Sept. 2011

13 departments, 26 faculty investors

Retired Oct. 2016

CARTER

10,368 cores

Installed April 2012 – Retired 2017

26 departments, 60 faculty investors

#54 on June 2012 Top 500

CONTE

9,280 Xeon cores (69,900 Xeon Phi)

Installed August 2013 – Retired 2018

26 departments, 68 faculty investors

#28 on June 2013 Top 500

DATA DEPOT

2.5 PB of disk storage
Installed Nov. 2014

550+ faculty investors from every academic college

RICE

13,200 cores, Installed May 2015

33 departments, 77 faculty investors

HALSTEAD

10,160 cores, Installed December 2016

39 departments, 79 faculty investors

BROWN

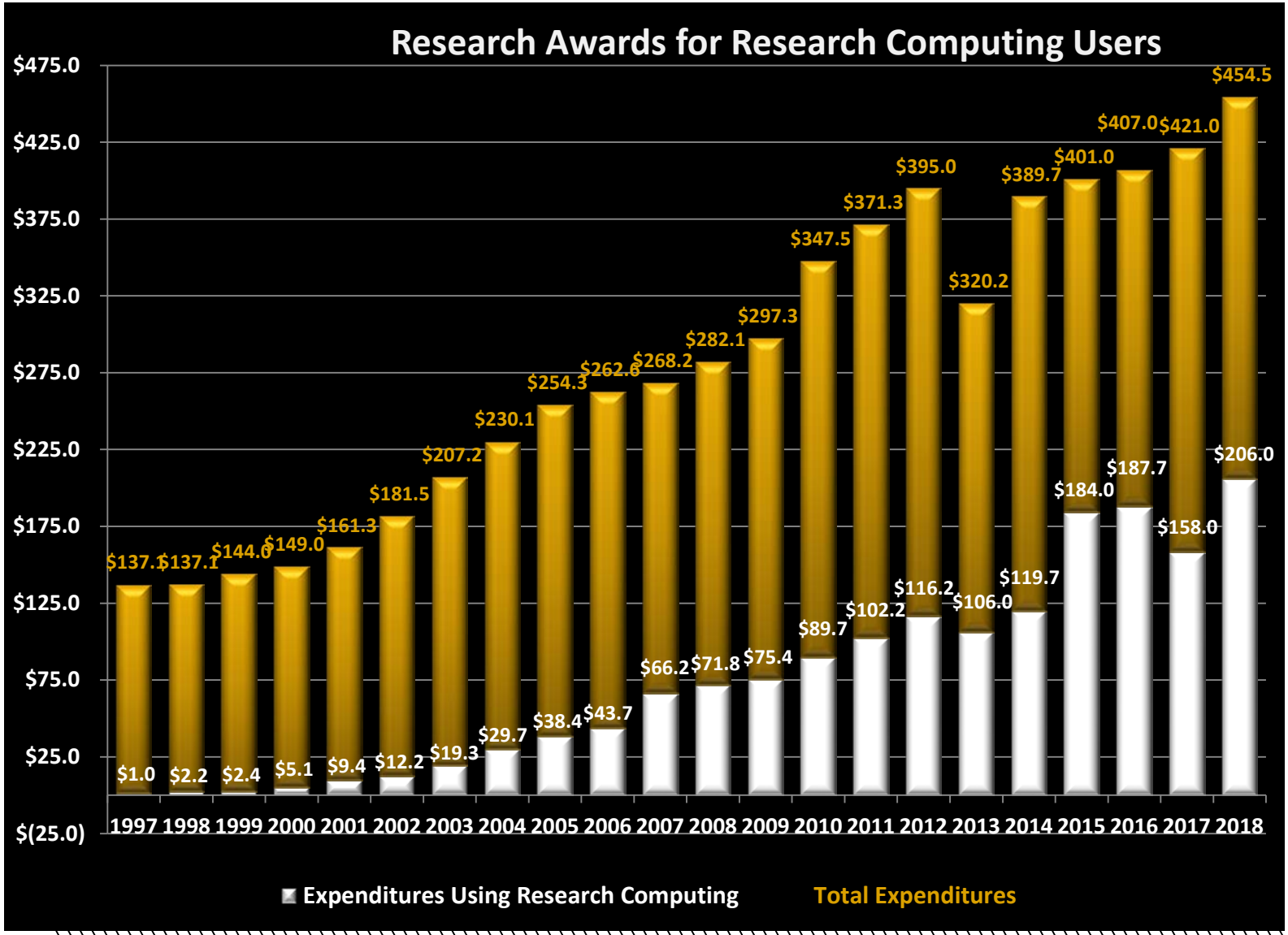
13,200 cores
Installed October 2017
36 departments
87 faculty investors

$4439 for 5 years of service

#302 on Nov 2017 Top 500


#105 STEELE 2008
#102 COATES 2009
#150 HANSEN 2010 (est.)
#126 ROSSMANN 2011
#54 CARTER 2012
#28 CONTE 2013
#166 RICE 2015
#302 BROWN 2017
PURDUE COMMUNITY CLUSTERS
TOP500 RANKINGS

Partner Productivity – Research Dollars



Scratch Storage for HPC

Community Cluster Storage

- Community Clusters are proven resources for computing, but require powerful storage systems to keep them fed
- Each system is built with a large, fast scratch filesystem tightly coupled with the compute nodes.
  - **Brown: 3.5PB @ 40 GB/sec, 400,000 IOPS**

- Short-term storage
- Per-user
- Large quotas (100-200 TB/user)
- High-speed, for active workflows
- Subject to purge after 60 days
- Not built for redundancy
  - No snapshots, backups

PURDUE UNIVERSITY

2018 Community Cluster

Prof. Lillian Moller Gilbreth

Gilbreth

- Based on faculty feedback:

  A GPU-based system ideal for **machine learning, AI, big data science** – as well as FEA, Chemistry, MD
- 50 nodes, 100 GPUs
  - ***1 PF of single-precision computing!***
- 2-3PB parallel filesystem storage
- Flash storage

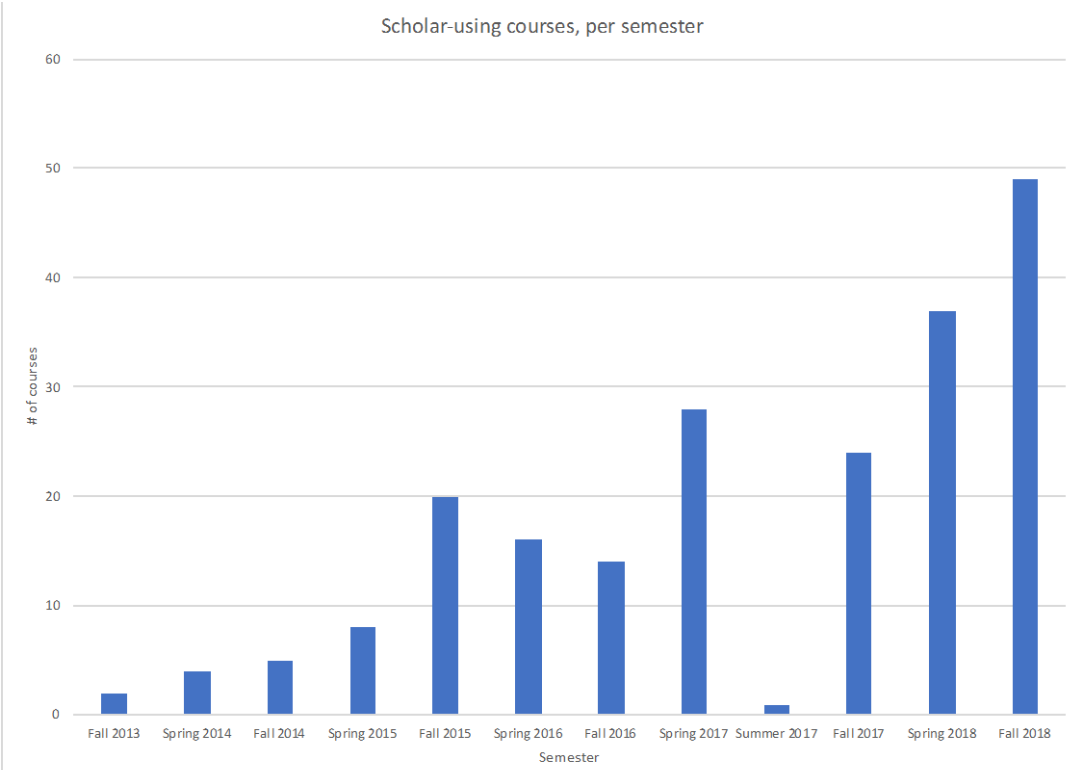
- Annual subscription fee for access

In Early Access Now!

PURDUE UNIVERSITY

Scholar – HPC and Data Science Education

Dedicated Cluster for HPC and Data Science



2035 students using Scholar to learn HPC and data science last fall semester!

2500 students in 36 classes already this Spring

Dept	# of Courses
AAE	6
ABE	2
AGRY	3
ANSC	1
BIOL	4
BME	1
CGT	1
CHM	5
CNIT	3
CS	6
EAPS	17
ECE	3
EEE	1
FS	1
HORT	1
IE	1
LIBR	1
ME	5
MGMT	9
NUTR	1
PHYS	1
STAT	9

PURDUE UNIVERSITY

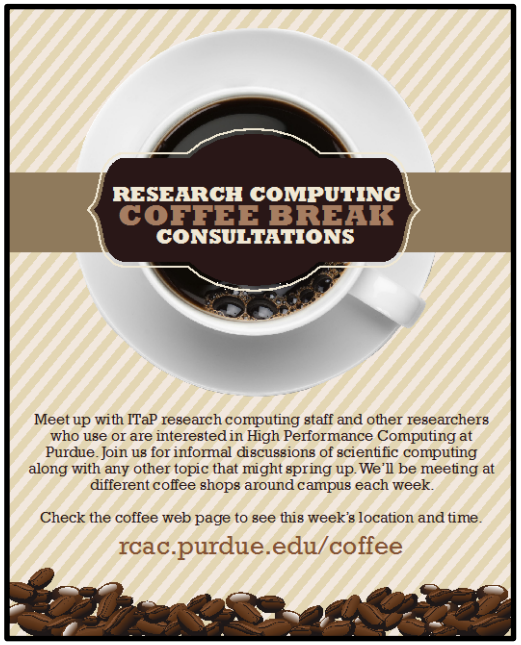
EXPERTISE

PURDUE UNIVERSITY

Education, Training, Expertise

Our key value – computational experts

- New faculty orientations
- One-on-one consultations
- UNIX, MPI, HPC, Big Data, Python, R training offerings
- Cyberinfrastructure seminars



software carpentry

ADVANCED DOMAIN EXPERTISE

Data Science
Chemistry
Physics
Astrophysics
Earth and Atmospheric Sciences
Computer Science
Chemical Engineering
Electrical and Computer Engineering
Cell and Molecular Biology
Entomology

PURDUE UNIVERSITY

THANK YOU

www.rcac.purdue.edu
psmith@purdue.edu