Research Computing Building Blocks

INFRASTRUCTURE FOR DATA AT PURDUE

PRESTON SMITH, DIRECTOR OF RESEARCH SERVICES

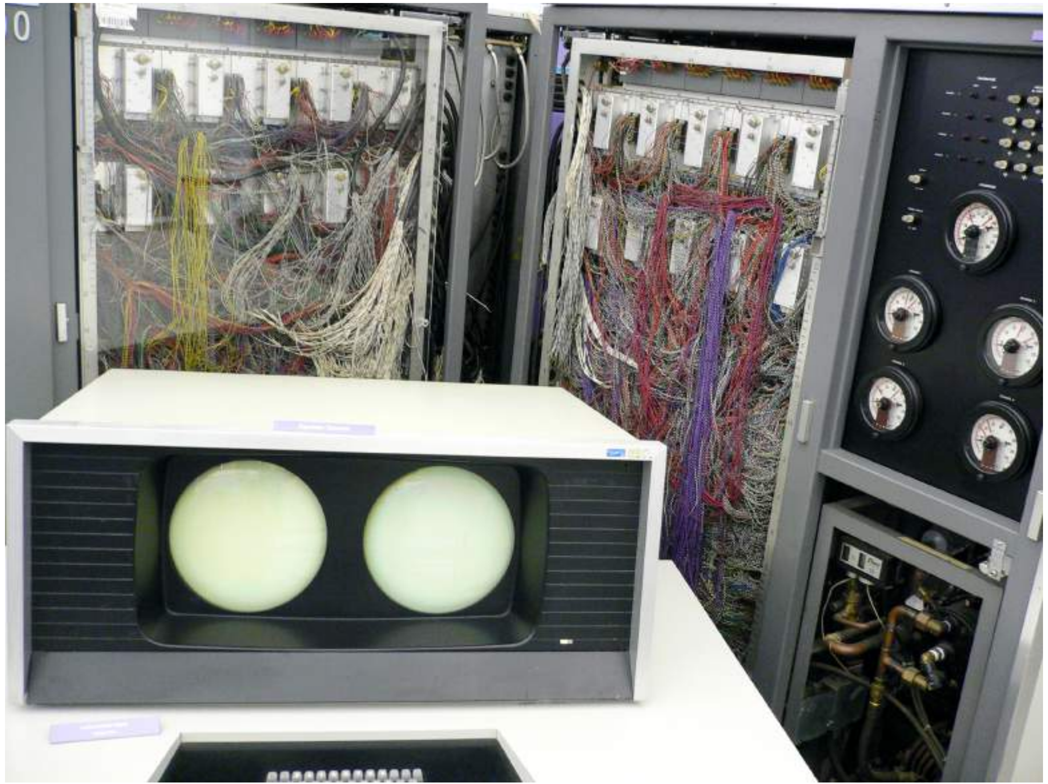
PSMITH@PURDUE.EDU

http://www.geartechnology.com/blog/wp-content/uploads/2015/11/opportunity-396265_640.jpg

Discussion

WHAT ARE THE GAPS BETWEEN THE BUILDING BLOCKS AND THE SCIENCE?

Data in IT



https://upload.wikimedia.org/wikipedia/commons/e/e3/CDC_6600_introduced_in_1964.jpg

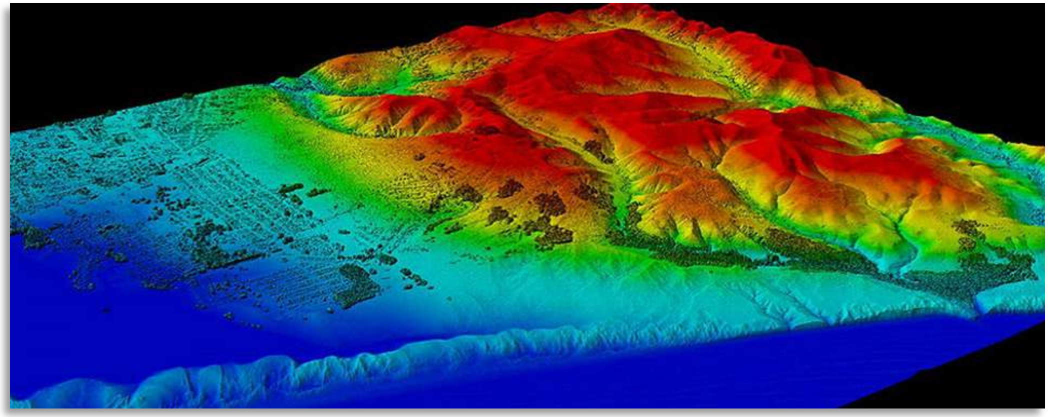
IT has always been about data! Computing and data are inextricably linked.

Purdue has had computing on campus for a very long time, since the days of the CDC 6500 in the 1960s.

We see both analysis and simulation!

What is big data?

- Not just Facebook-style analytics!
- 3.5 PB of high-energy physics detector data
- 1 PB of climate model data
  - 90 TB in an active workflow!
- 200 TB of astrophysics simulations
- 150 TB of CFD model output
- 120 TB of audio files
- 100 TB of actively-used next-gen sequencing data
  - Millions of files used in an active workflow
- 10s of TB of video files
- 5 TB of electron microscope images generated per day
- ..to the 75% of users on Conte using less than 1TB
- … and to the social science researcher with stacks of excel sheets



Big data: A data set that is larger/faster/more complex than one feels comfortable dealing with.

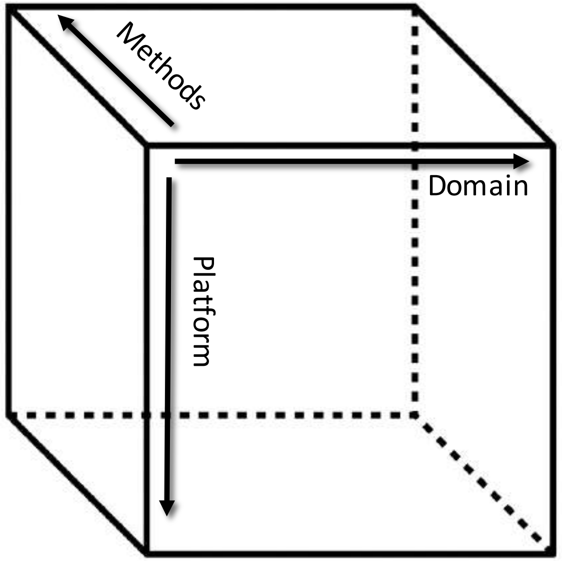
Scope of Data problems at Purdue

	Domain 1	Domain 2	Domain 3
Platform 1			
Platform 2			
Platform 3			
Platform 4			

X

Not just a matrix

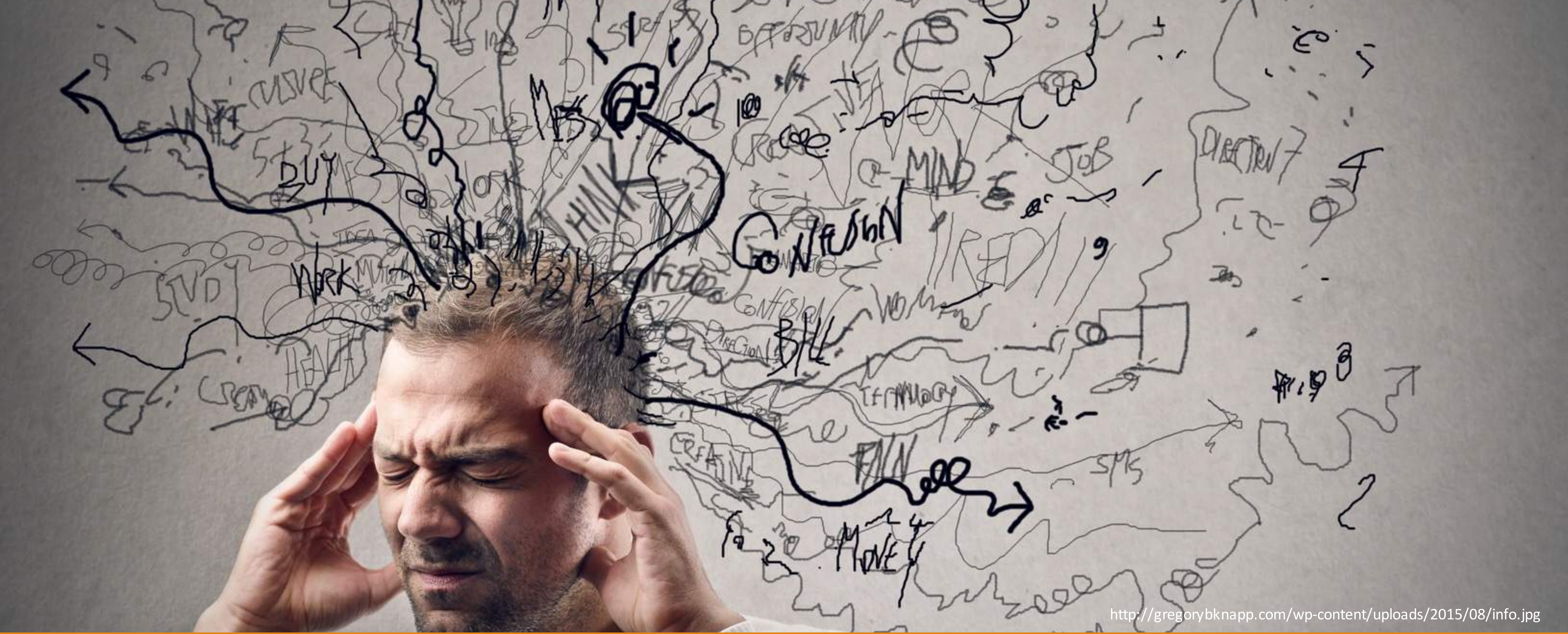
Scope of Data problems at Purdue



A 3D cube of:

- Domain
- Technology/Methods
- Computing Platform

Biointormatics - using Bioconductor on the Snyder Supercomputer



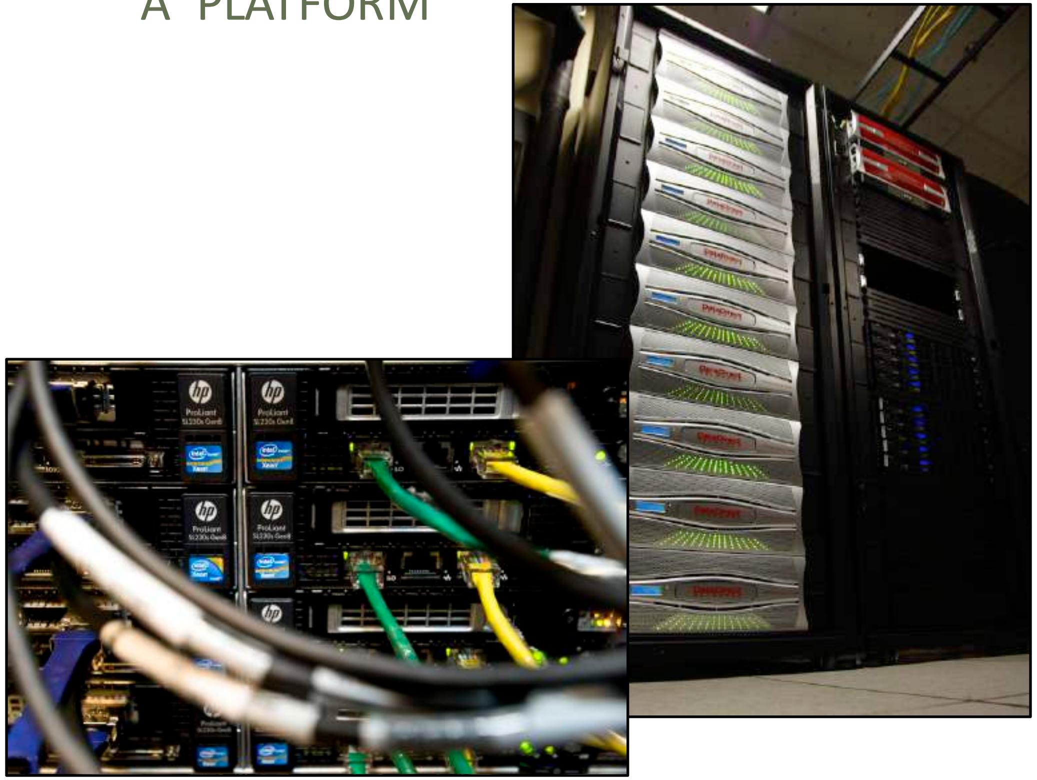
http://gregorybknapp.com/wp-content/uploads/2015/08/info.jpg

Discussion: How can we scope this challenge?

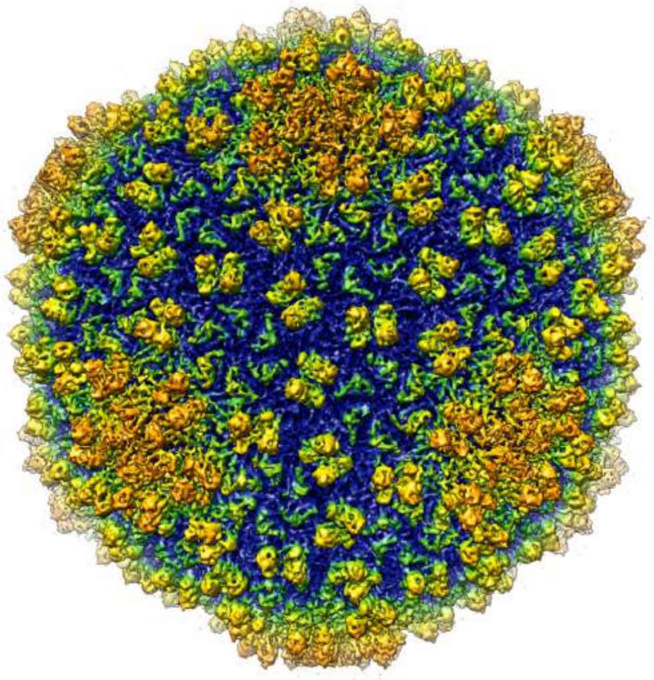
Can there be a one-stop place to go?

Research Computing Support of Data

A PLATFORM



VARIOUS DOMAINS AND APPLICATIONS



https://news.uns.purdue.edu/images/+2008/jiang-bacteriophage.jpg

Our Domains

DOMAINS

Chemistry

Physics

Astrophysics

Earth and Atmospheric Sciences

Computer Science

Chemical Engineering

Electrical and Computer Engineering

Cell and Molecular Biology

Agriculture

APPLICATION SPACES

Molecular Dynamics

Image Processing

Quantum Chemistry

Weather Modeling

Machine Learning

Big Data

Computer Architecture

Finite Element Analysis

Statistics

Bioinformatics

Geospatial

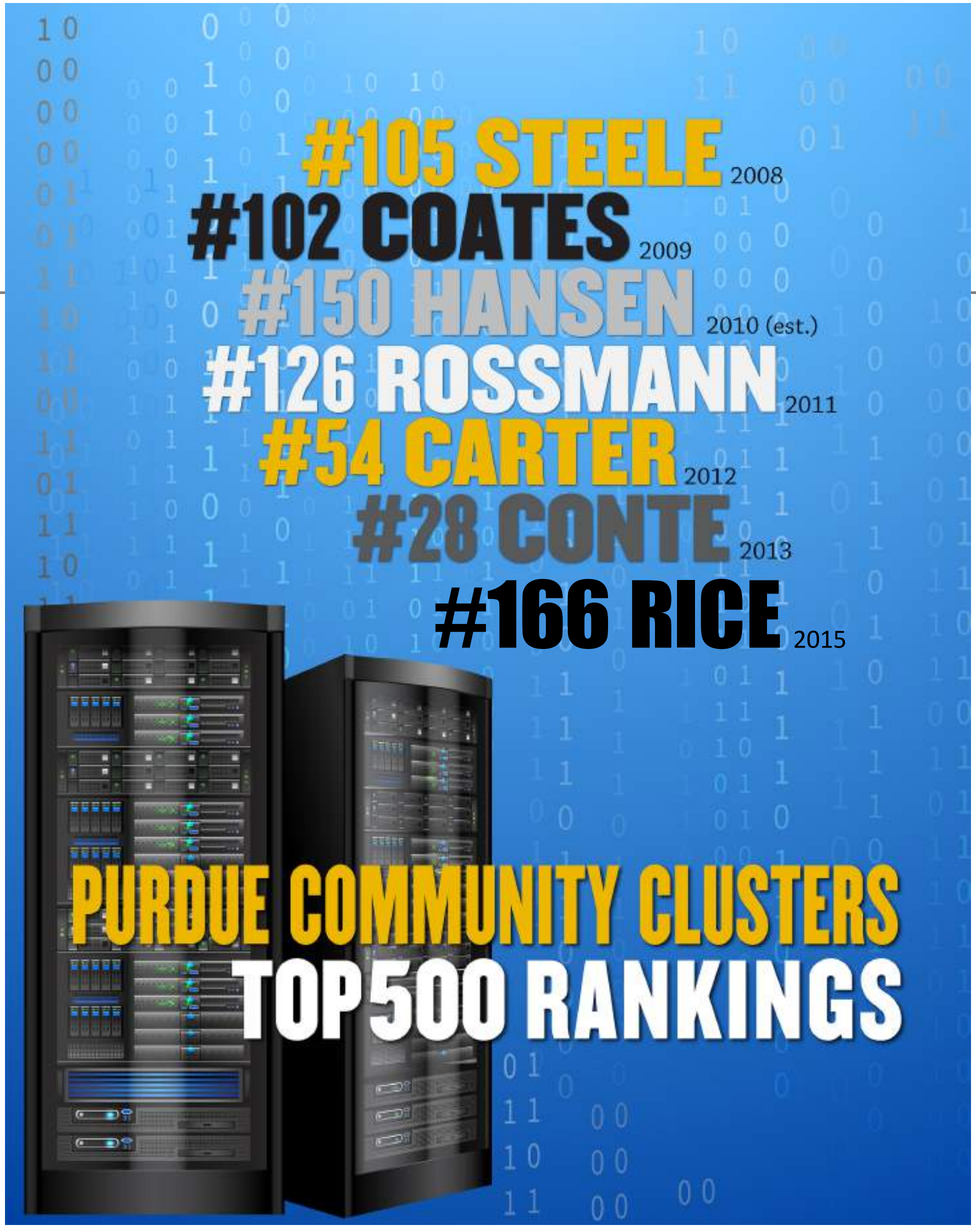
Remote Sensing

Visualization

Community Cluster Program

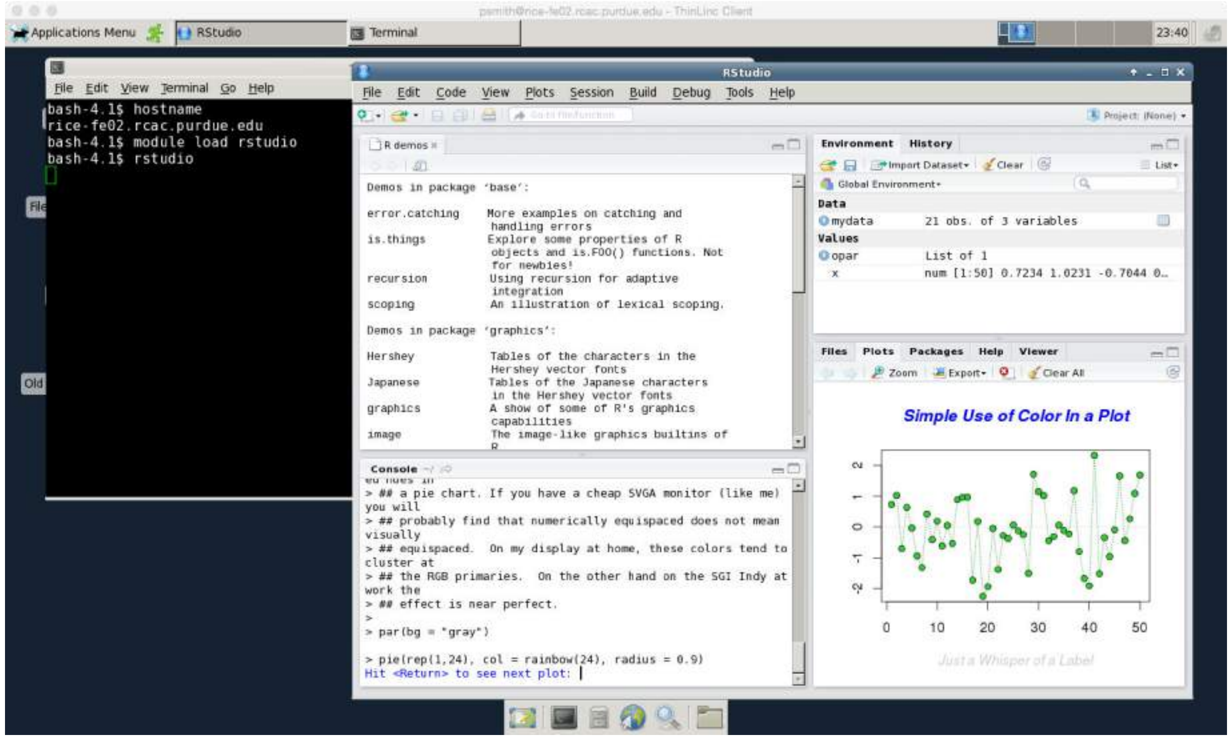
2015 Systems:

Rice – Parallel Computing
Snyder – Data-Intensive Life Science
Hammer – High-Throughput Computing



Steele Cluster, 2008

Your Personal Supercomputer



psmith@rice-fe02.rcac.purdue.edu - ThinLinc Client
Applications Menu
RStudio
Terminal
23:40
File Edit View Terminal Go Help
RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
R demos
Environment History
Import Dataset
Clear
List
Global Environment
Files Plots Packages Help Viewer
Zoom
Export
Clear All
Console
Simple Use of Color In a Plot
2
1
0
-1
-2
0
10
20
30
40
50
Just a Whisper of a Label

```
bash-4.1$ hostname
rice-fe02.rcac.purdue.edu
bash-4.1$ module load rstudio
bash-4.1$ rstudio
```

```
Demos in package 'base':

error.catching     More examples on catching and
                    handling errors
is.things          Explore some properties of R
                    objects and is.FOO() functions. Not
                    for newbies!
recursion          Using recursion for adaptive
                    integration
scoping            An illustration of lexical scoping.

Demos in package 'graphics':

Hershey            Tables of the characters in the
                    Hershey vector fonts
Japanese           Tables of the Japanese characters
                    in the Hershey vector fonts
graphics           A show of some of R's graphics
                    capabilities
image              The image-like graphics builtins of
                    R
```

```
ed hues in
> ## a pie chart. If you have a cheap SVGA monitor (like me)
you will
> ## probably find that numerically equispaced does not mean
visually
> ## equispaced.  On my display at home, these colors tend to
cluster at
> ## the RGB primaries.  On the other hand on the SGI Indy at
work the
> ## effect is near perfect.
>
> par(bg = "gray")

> pie(rep(1,24), col = rainbow(24), radius = 0.9)
Hit <Return> to see next plot: 
```

```
Data
mydata        21 obs. of 3 variables
Values
opar          List of 1
  x           num [1:50] 0.7234 1.0231 -0.7044 0…
```

Commonly-used software, toolkits, compilers, and libraries installed and maintained by ITaP computational scientists.

Easy-to-use graphical access available.

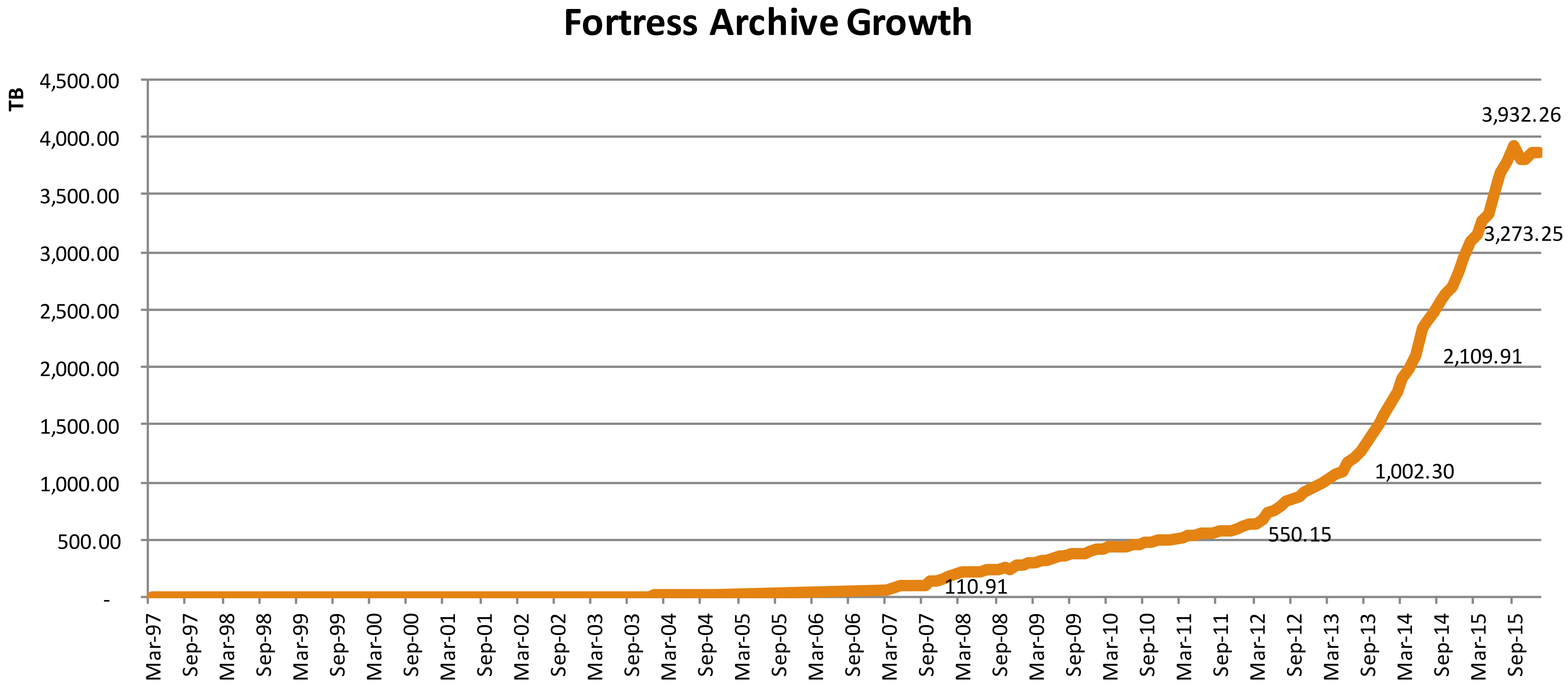
Data Storage

https://upload.wikimedia.org/wikipedia/commons/e/e7/Interior_of_StorageTek_tape_library_at_NERSC_(1).jpg

The Fortress archive is a large, long-term, multi-tiered file caching and storage system utilizing both online disk and robotic tape drives.

Ideal for permanent storage of your research data.

Explosions of Data



Data Storage

The Research Data Depot is a high-capacity, high-performance, reliable and secure data storage service designed, configured and operated for a lab's active research data.

220 research labs

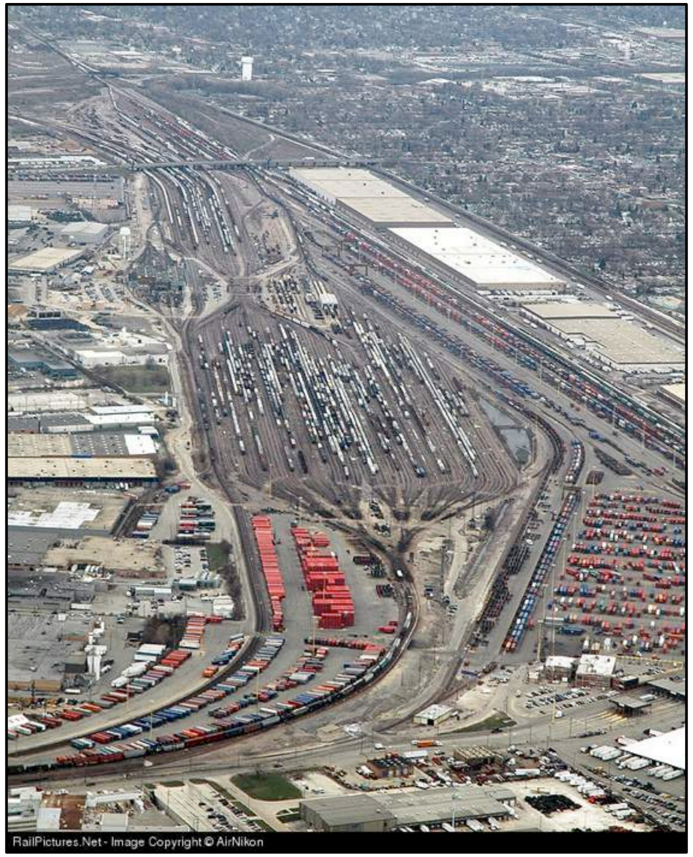
.75 PB allocated

More than just file services!

MySQL

PostgreSQL

mongoDB



RailPictures.Net - Image Copyright © AirNikon

Data Storage

Supercomputer systems are built with a 1 Petabyte+ scratch filesystem for running jobs.

Holding input data, writing results.
Data copied to Fortress or Data Depot.

100T allocated per user.

Very high-speed, very scalable.

No data protection beyond RAID!

http://michaelroud.com/wp-content/uploads/2013/11/Cost-of-Being-an-Actor-e1384976639996.jpg

Discussion: What about the costs of data storage?

At large scale, costs add up quickly when borne by the researcher.

Data Analytics

http://bit.ly/1QCennM

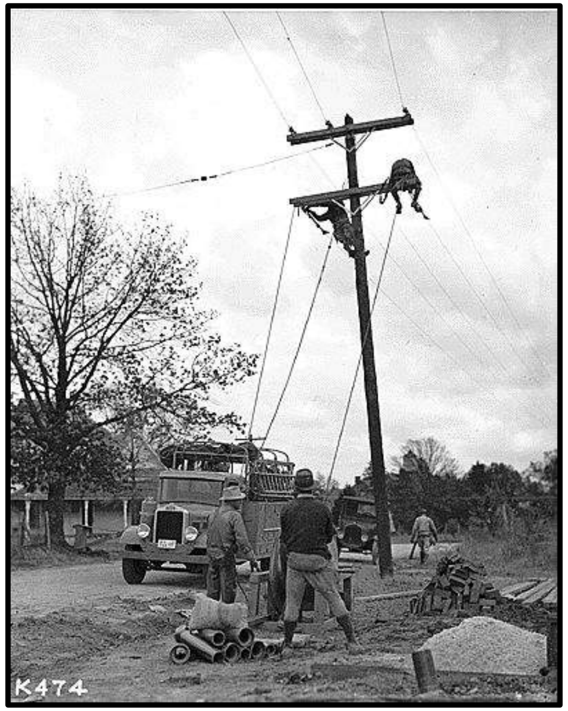
- “hathi” Hadoop cluster for prototyping big data applications
- Spark, Hbase, Hive, Pig, Storm etc.

Spark Software fully supported on community clusters as well!

Research Networking

As science gets more data-intensive – researchers require increasing amounts of bandwidth

The last mile to the labs is key!



https://www.nwcouncil.org/media/24501/rural.jpg

https://pmcdeadline2.files.wordpress.com/2014/05/greenacres132__140501163754.jpgrur

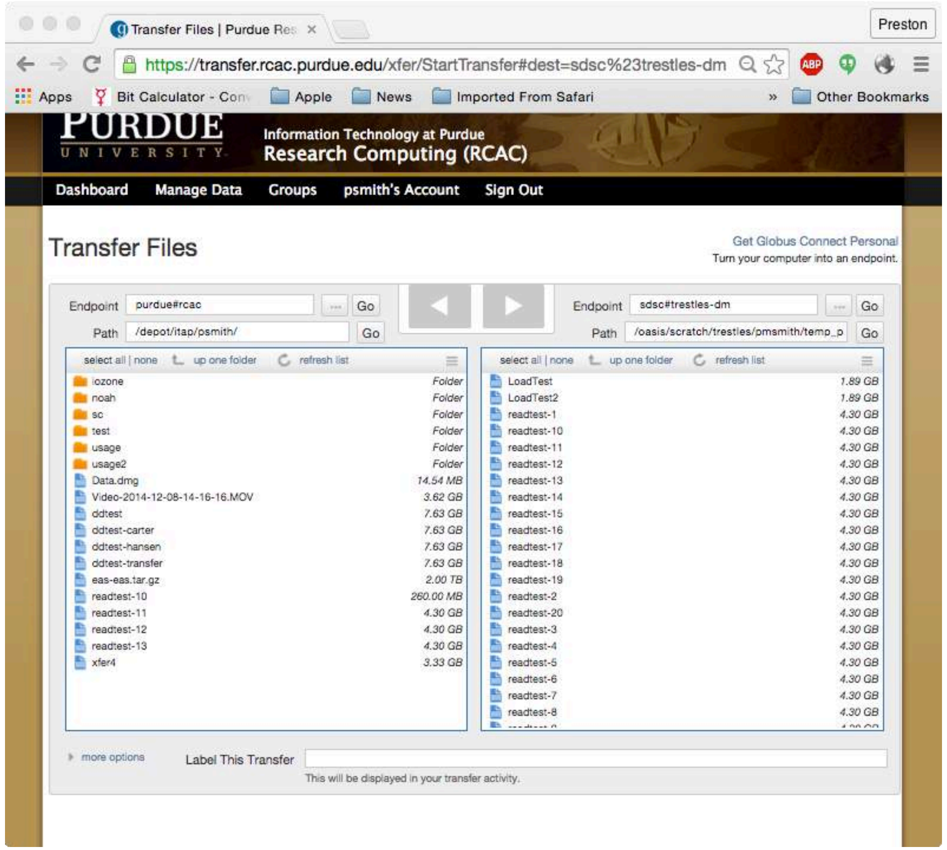
Instruments

https://upload.wikimedia.org/wikipedia/commons/7/7b/Illumina_MiSeq_sequencer.jpg

Instruments are getting cheaper, more common, and generate more data.

High-speed (10Gb+) connections for labs and instruments to move data into clusters, storage, and research WAN connections.

Data Transfer and Sharing



https://transfer.rcac.purdue.edu/xfer/StartTransfer#dest=sdsc%23trestles-dm

Transfer Files

Endpoint: purdue#rcac
Path: /depot/itap/psmith/

iozone	Folder
noah	Folder
sc	Folder
test	Folder
usage	Folder
usage2	Folder
Data.dmg	14.54 MB
Video-2014-12-08-14-16-16.MOV	3.62 GB
ddtest	7.63 GB
ddtest-carter	7.63 GB
ddtest-hansen	7.63 GB
ddtest-transfer	7.63 GB
eas-eas.tar.gz	2.00 TB
readtest-10	260.00 MB
readtest-11	4.30 GB
readtest-12	4.30 GB
readtest-13	4.30 GB
xfer4	3.33 GB

Endpoint: sdsc#trestles-dm
Path: /oasis/scratch/trestles/pmsmith/temp_p

LoadTest	1.89 GB
LoadTest2	1.89 GB
readtest-1	4.30 GB
readtest-10	4.30 GB
readtest-11	4.30 GB
readtest-12	4.30 GB
readtest-13	4.30 GB
readtest-14	4.30 GB
readtest-15	4.30 GB
readtest-16	4.30 GB
readtest-17	4.30 GB
readtest-18	4.30 GB
readtest-19	4.30 GB
readtest-2	4.30 GB
readtest-20	4.30 GB
readtest-3	4.30 GB
readtest-4	4.30 GB
readtest-5	4.30 GB
readtest-6	4.30 GB
readtest-7	4.30 GB
readtest-8	4.30 GB

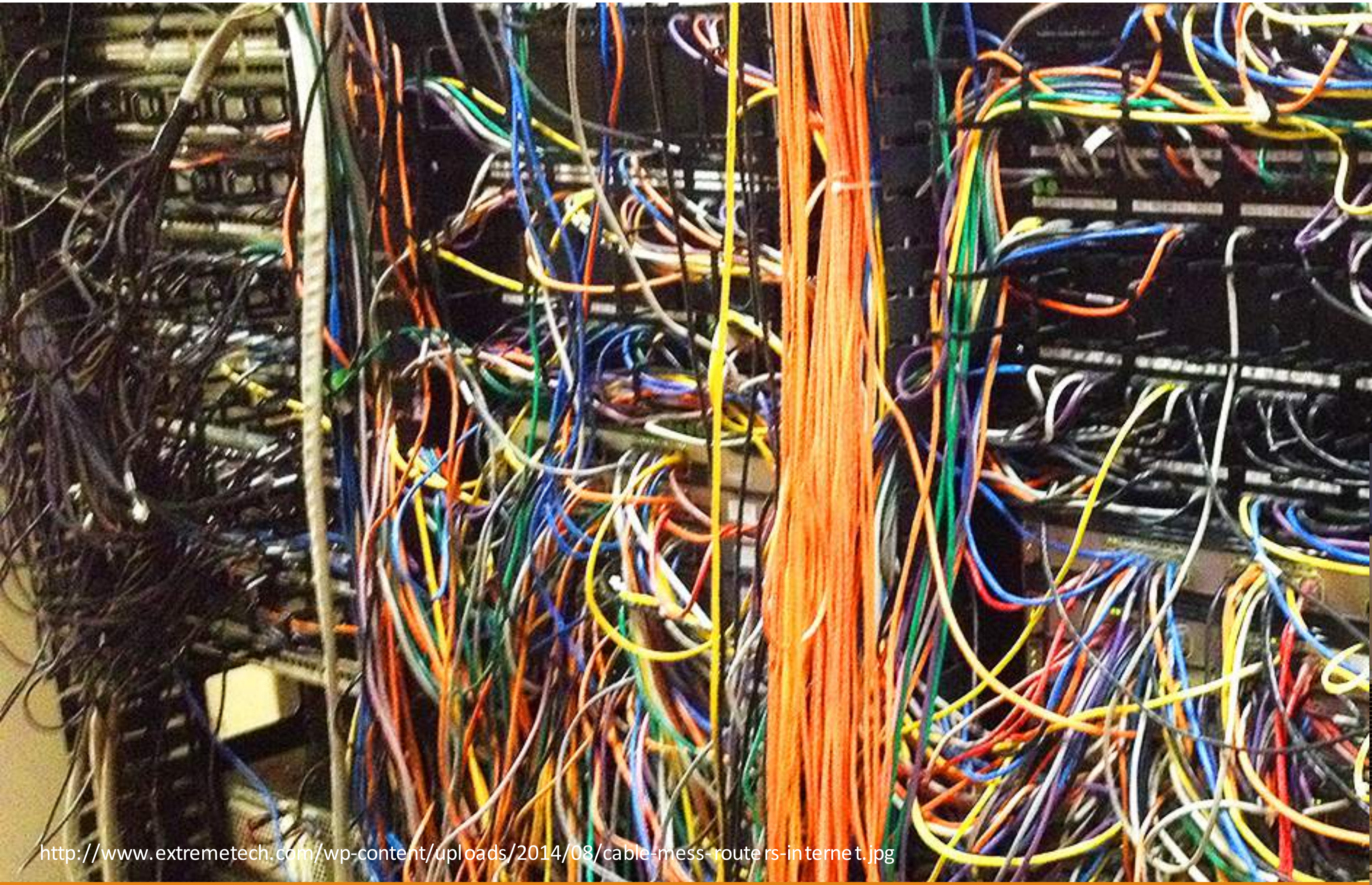
Label This Transfer:

This will be displayed in your transfer activity.

Transfer and share large datasets….

…. With dropbox-like characteristics ….

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



http://www.extremetech.com/wp-content/uploads/2014/08/cable-mess-routers-internet.jpg

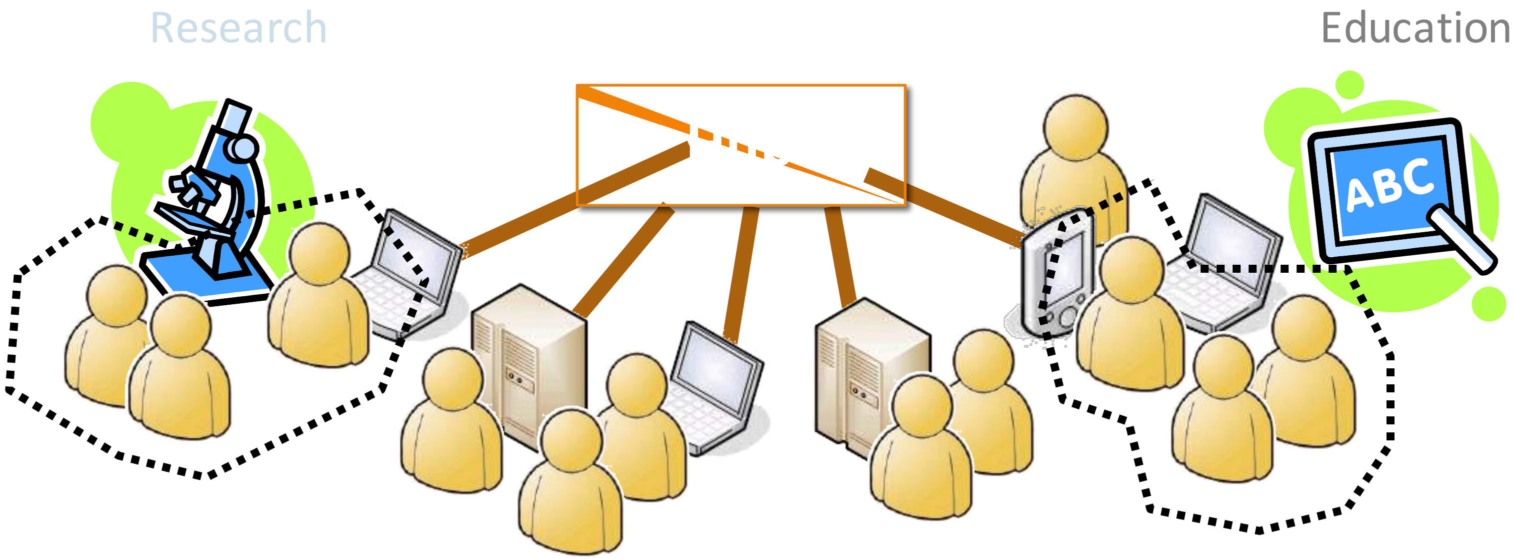
Networking

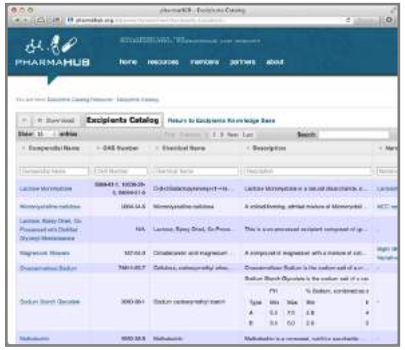
How to balance security, performance, and accessibility to have a high-speed, friction-free end-to-end experience between the lab and HPC?

Instruments

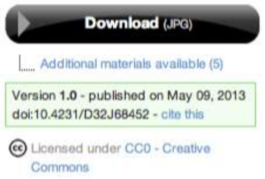
How can we reliably collect and move data?

Hubzero: Collaboration, Online Simulation, and Data





- ✓ Databases and digital publications
- ✓ Uploaded by researchers in the community
- ✓ Digital Object Identifiers and license options
- ✓ Data ↔ tools for analysis


Download (JPG)

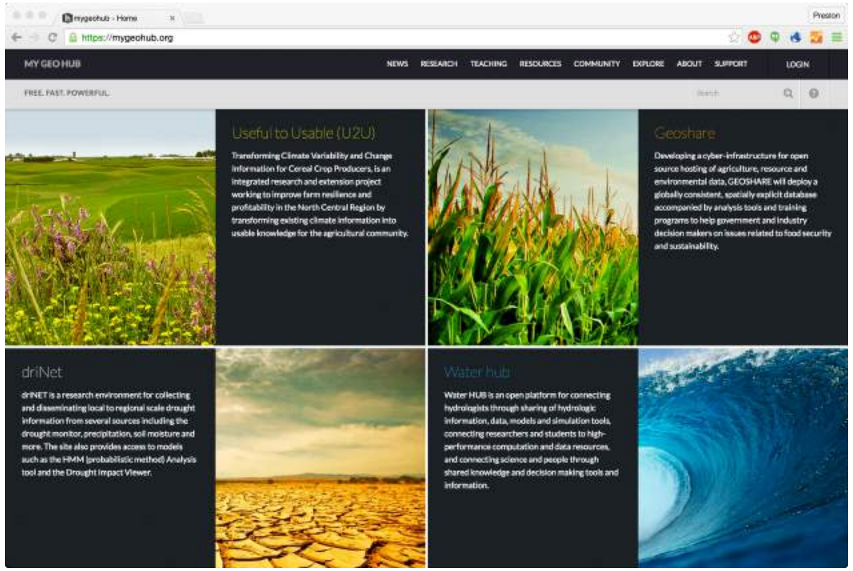
Additional materials available (5)

Version 1.0 - published on May 09, 2013
doi:10.4231/D32J68452 - cite this

Licensed under CC0 - Creative Commons

Science Gateways

Web-based portals that enable a community to share data, tools, and collaborate.



Useful to Usable (U2U)

Transforming Climate Variability and Change Information for Cereal Crop Producers, is an integrated research and extension project working to improve farm resilience and profitability in the North Central Region by transforming existing climate information into usable knowledge for the agricultural community.

Geoshare

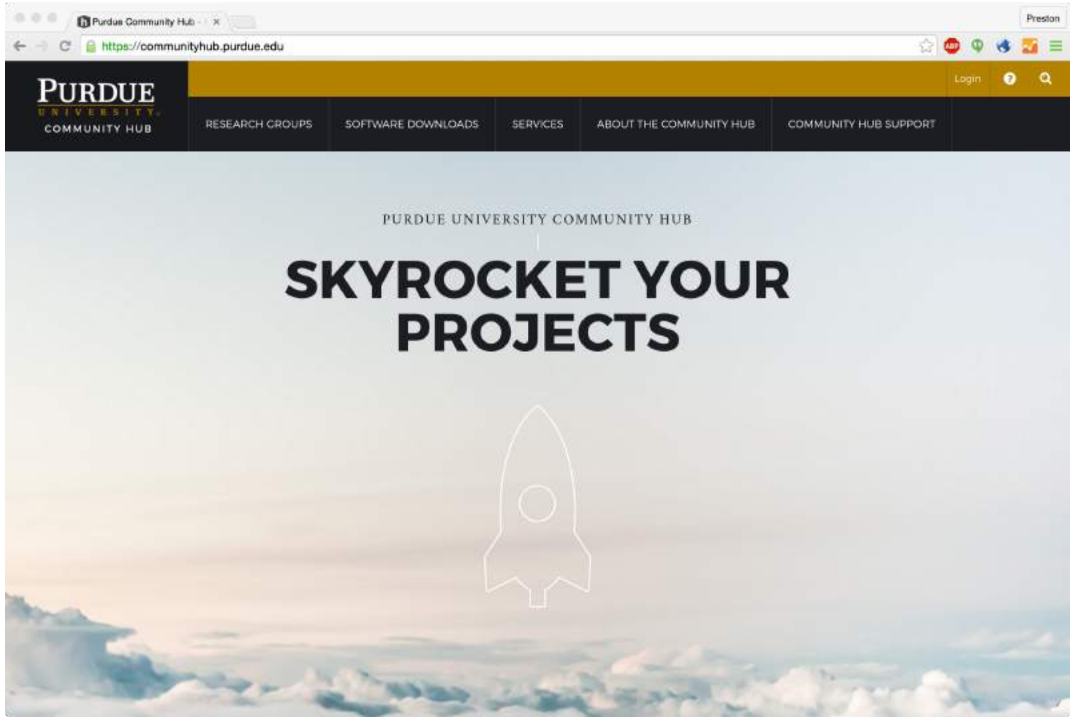
Developing a cyber-infrastructure for open source hosting of agriculture, resource and environmental data, GEOSHARE will deploy a globally consistent, spatially explicit database accompanied by analysis tools and training programs to help government and industry decision makers on issues related to food security and sustainability.

driNet

driNET is a research environment for collecting and disseminating local to regional scale drought information from several sources including the drought monitor, precipitation, soil moisture and more. The site also provides access to models such as the HMM (probabilistic method) Analysis tool and the Drought Impact Viewer.

Water hub

Water HUB is an open platform for connecting hydrologists through sharing of hydrologic information, data, models and simulation tools, connecting researchers and students to high-performance computation and data resources, and connecting science and people through shared knowledge and decision making tools and information.



PURDUE UNIVERSITY COMMUNITY HUB

SKYROCKET YOUR PROJECTS

Research Solutions

A staffing gap exists between the science and the expertise in advanced research technology, for creating new solutions.

- Applied technology and software developers

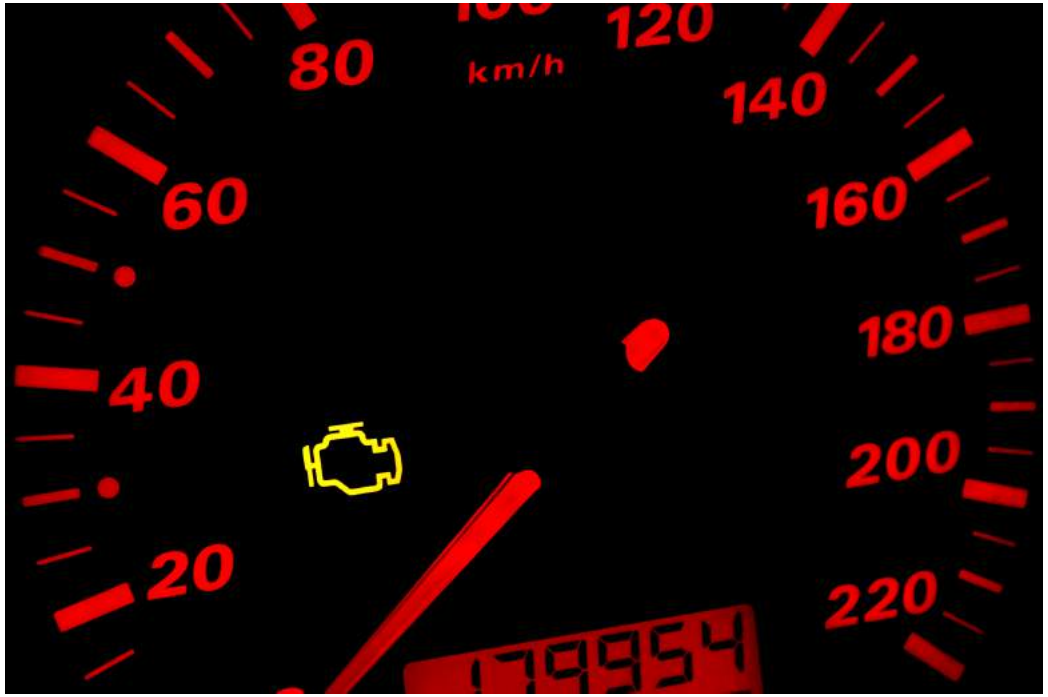
Computing Literacy

Our computational scientists are investing heavily in teaching faculty and students

- UNIX literacy
- Effective use of clusters
- Programming models (MPI)
- Visualization
- “Big Data” Tools
- Software carpentry

One-on-one instruction as well!

Computing Literacy



Is computing like a car?

As a driver going back and forth to campus, I could say “I don’t know how it does what it does, I just drive it”. It tells me when something goes wrong.

Should researchers be shielded from the details of how computing works for them?

Computing Literacy

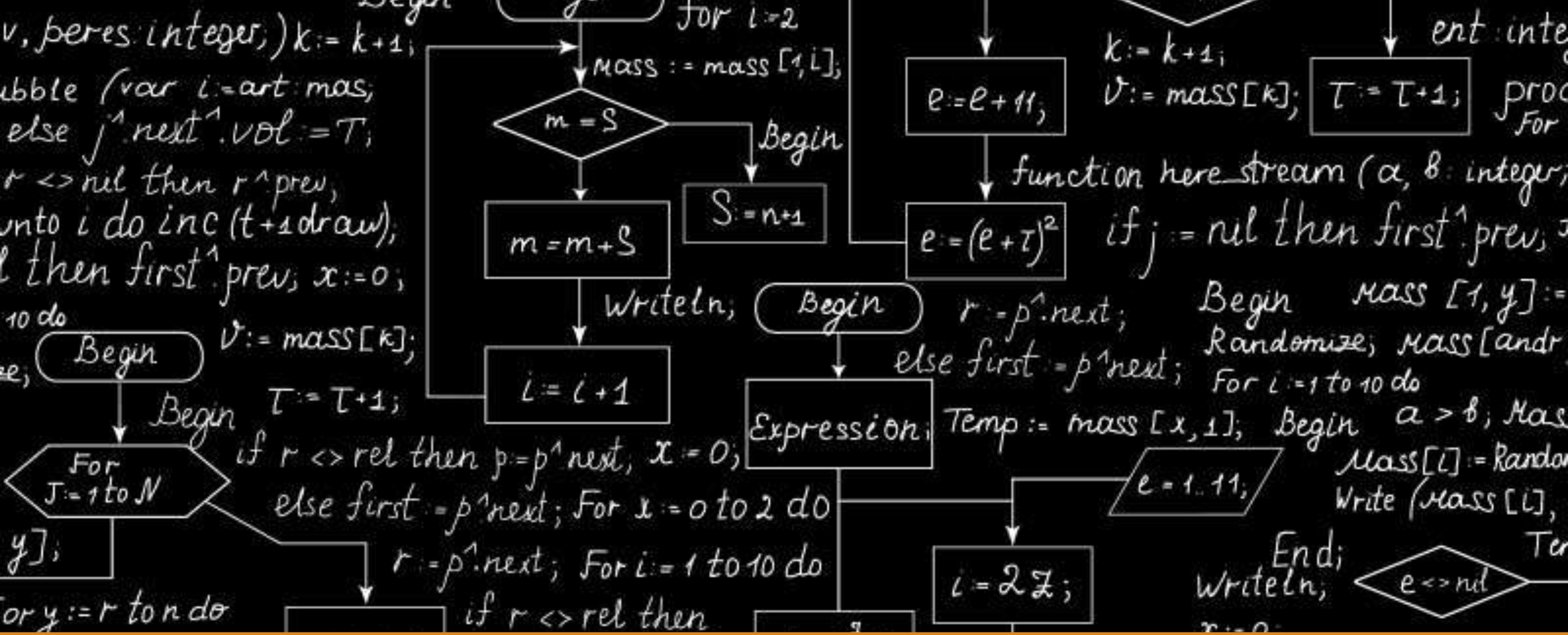
Or..

Is the driver in Indianapolis a better analogy?

There are people who make sure the track is in good shape and the car is running fast, but you can bet that the driver understands his car.

Downforce
Wind
Traffic

Heat
Tire Wear
Aerodynamics



Education

How do we train our graduate students to use the computing and data resources they need to develop into computationally-literate scientists?