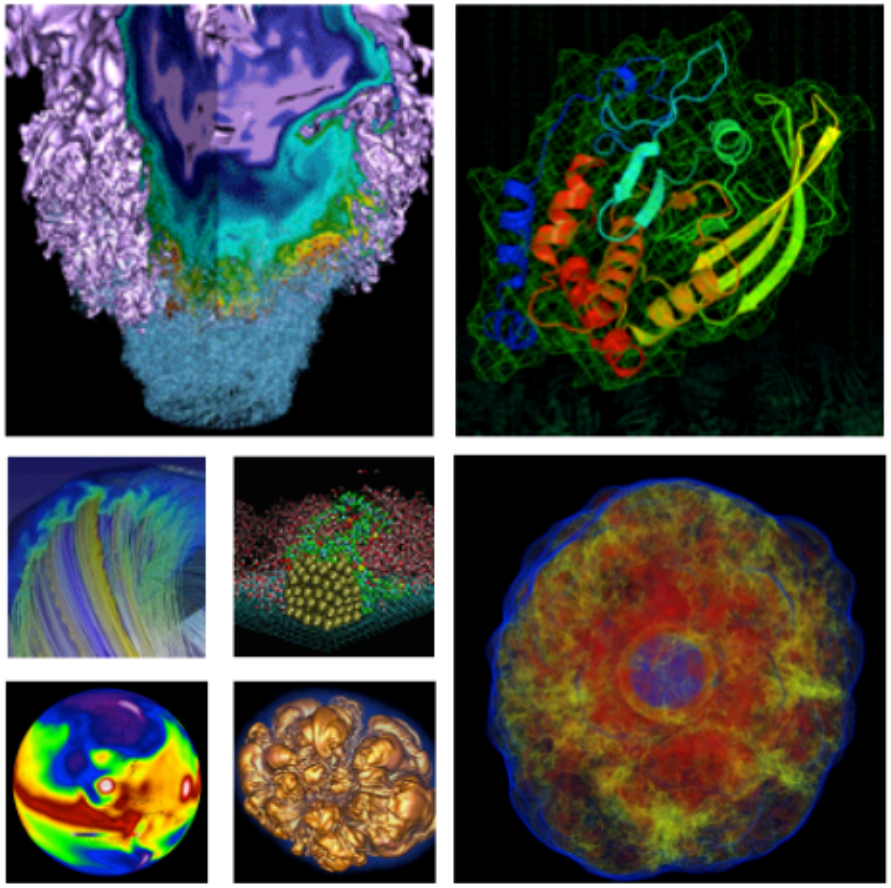
Data Analytics at NERSC



Rollin Thomas
rcthomas@lbl.gov
NERSC Data and Analytics Services

March 21, 2016
NERSC User Group Meeting

U.S. DEPARTMENT OF ENERGY | Office of Science

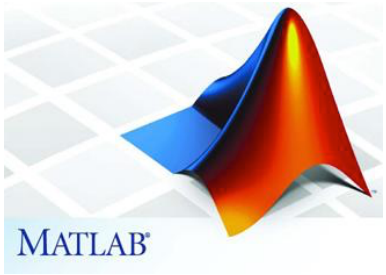
BERKELEY LAB

Introduction

- **Data Analytics**: The key to unlocking insight from massive and complex data sets.
- NERSC supports a variety of general-purpose analytics tools and services.
- This talk will cover:
  - Data analytics tools available on the Cray machines.
  - Other analytics services enabled through the web.
  - How to get help with data analytics at NERSC.
  - What's coming?

Data Analytics at NERSC

Data Analytics Tools



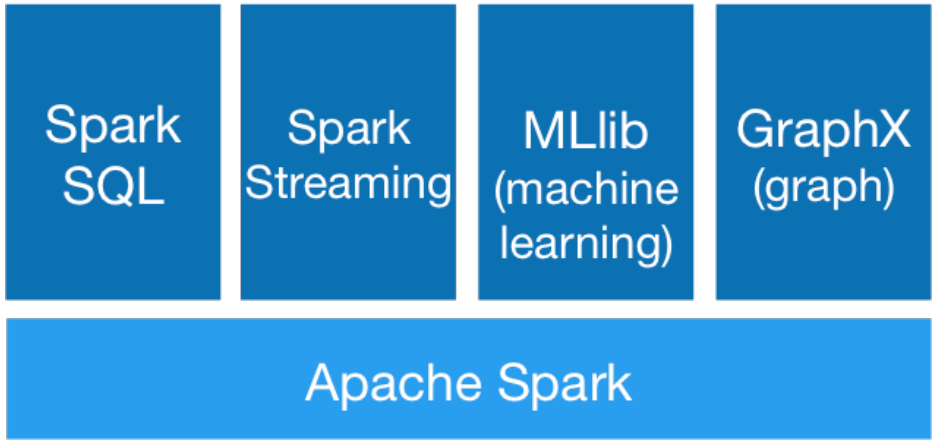
- **R:** Extensible language and environment for statistical computing and graphics.
- Linear, non-linear modeling, classical statistics, time series analysis, classification, clustering, visualization.
- To use R on Edison or Cori:
  - `module load R`
    - Interactive via login or compute nodes (salloc).
    - Or via batch script (sbatch).
    - Variety of approaches for achieving parallelism.
- Users may install packages in $HOME or ask for system-wide installation via consult@nersc.gov.

- **Python:** Interpreted, general-purpose, high-level programming language. Python 2.7.x and 3.4.x.
- Number of scientific computing packages: numpy, scipy, matplotlib, scikit-learn, mpi4py, ...
- To use Python on Edison or Cori, *always* module load:
  - `module load python` (NERSC-built)
    - Or, e.g., `module load python/2.7-anaconda`
    - Login, interactive (salloc), and batch (sbatch).
    - Parallelism: mpi4py, multiprocessing, Intel MKL.
- Users may install packages via pip, virtualenv, conda or ask for system-wide installation via consult@nersc.gov.

- Demand for Python at NERSC is large and increasing.
- Want Python to become more of a first-class citizen.
- Parallelism issues and progress:
  - Improved launch times by mounting /usr/common read-only with client-side caching on compute.
  - Greater awareness of tools and strategies for scaling up Python applications.
- (Too?) many choices in distribution space:
  - But certain Cray-specific subtleties (parallelism) require NERSC to build certain packages.
  - Anaconda Python now includes Intel MKL support.
  - Users are encouraged to consider Anaconda.

- **Spark:** Fast general purpose engine for large scale data processing, map-reduce, etc.

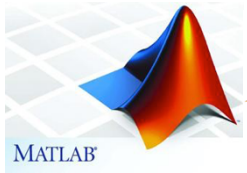
Computation Type	Spark Implementation
Machine Learning	MLib, Spark ML
Graph Computations	GraphX
Database Operations	Spark SQL
Streaming Analysis	Spark Streaming
Your Own Custom Analysis	Using Spark's Built In Functions



Computation types can be combined seamlessly all in the same piece of code.

- Spark has APIs in Java, Scala, Python, and R.
- NERSC recommends using Spark on Cori:
  - Large memory and I/O bandwidth requirements.
  - `module load spark`
    - Interactive (salloc) and batch (sbatch) supported.
    - Do not load the module until batch job launches!
- Spark support is experimental.
- Contact consult@nersc.gov if you have questions.

- Also in the ML space: **neon**, a Python-based, scalable Deep Learning executable and library.




- Tools requiring a license checkout:
  - **Matlab:** `module load matlab`
    Compute, visualize, and program in a familiar environment that "looks like math."
  - **Mathematica:** `module load mathematica`
    Symbolic mathematics, numerical calculations, visualization in a notebook interface.
  - **IDL:** `module load idl`
    Interactive data analysis and visualization environment.

- More tools (no license checkout):
  - **ROOT:** `module load root`
    Object-oriented framework for large-scale data analysis. Particle physics to data mining.
  - **Julia:** `module load julia` (Cori)
    Experimental high-level language for scientific computing with powerful type semantics.
- A NERSC best practice:
  - **Use NX** for interactive visualization tools and Mathematica notebooks.
  - http://www.nersc.gov/users/connecting-to-nersc/using-nx/

Web-Enabled Data Analytics Tools

Jupyter/IPython and RStudio

IP[y]: IPython
Interactive Computing

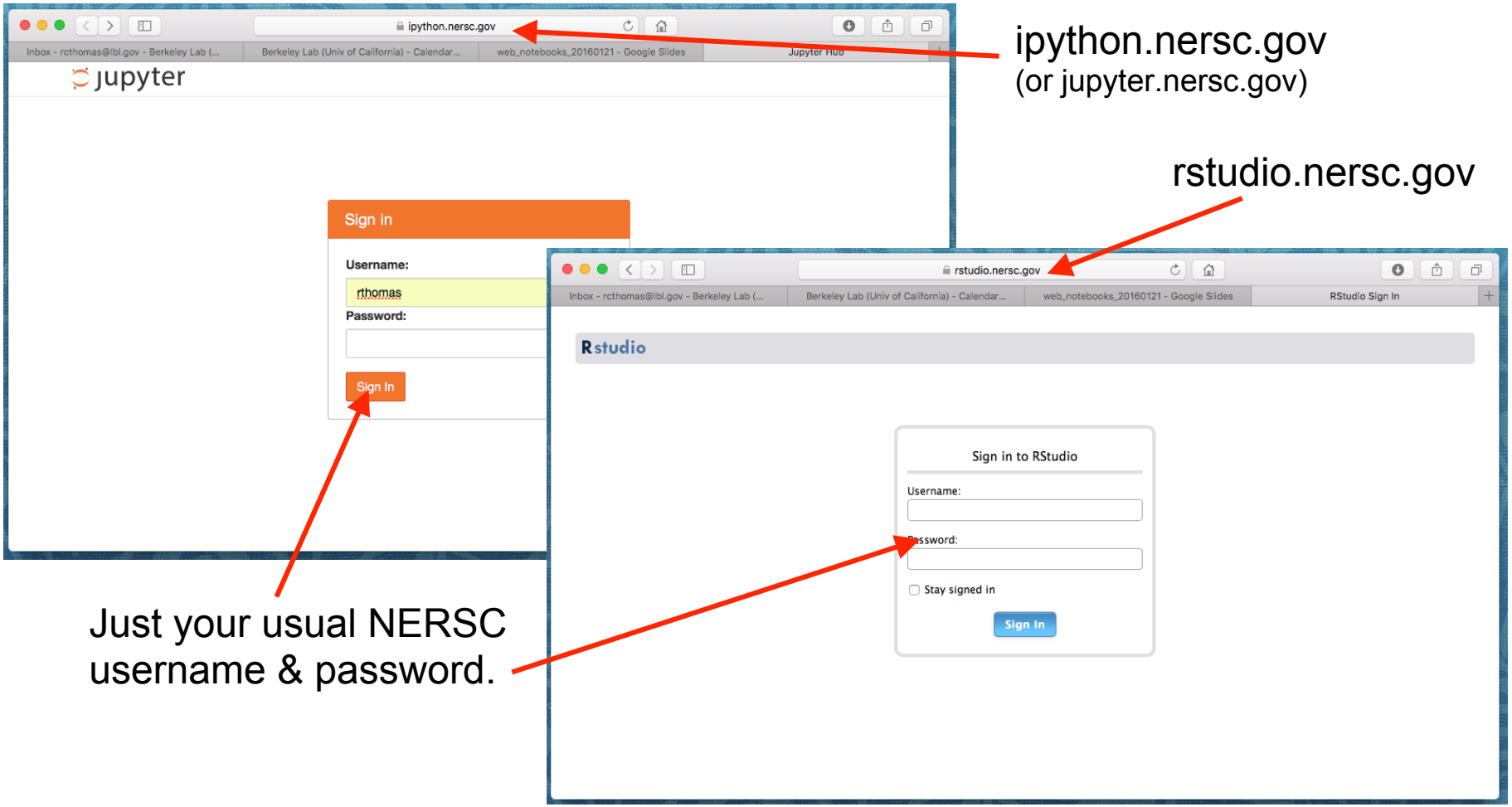
Powerful interactive shell originally developed for Python.
Also provides a web browser-based **notebook** supporting:

- Execution of code and annotation with text.
- In-line plotting and visualization.
- Interactive widgets.

Jupyter is the notebook part (language agnostic).
IPython is the Python shell and a Jupyter "kernel."

Integrated development environment (IDE) for R.
(R is also available at NERSC at the command line.)
RStudio provides a web browser-based IDE.

How Do I Work This?

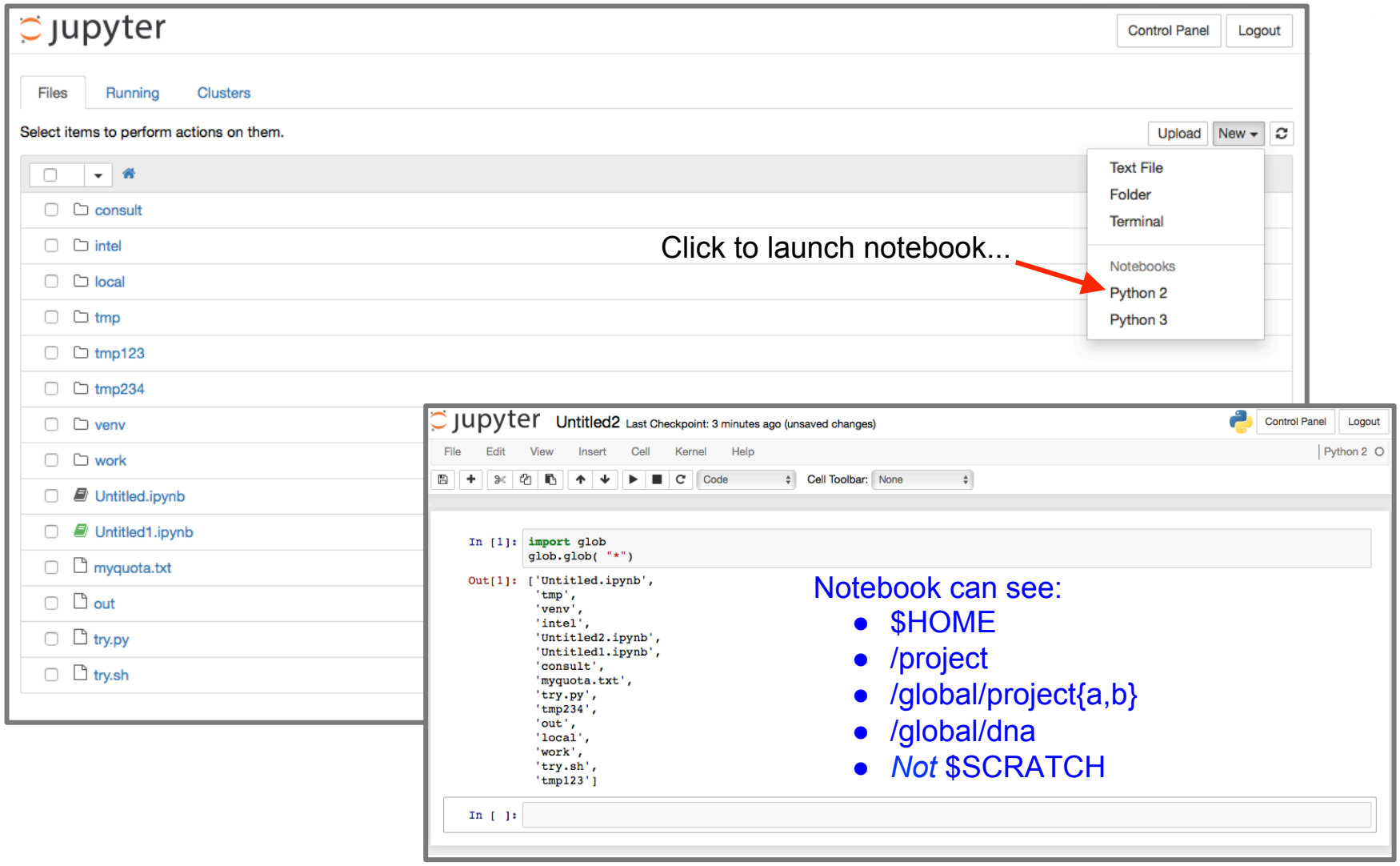


ipython.nersc.gov
(or jupyter.nersc.gov)

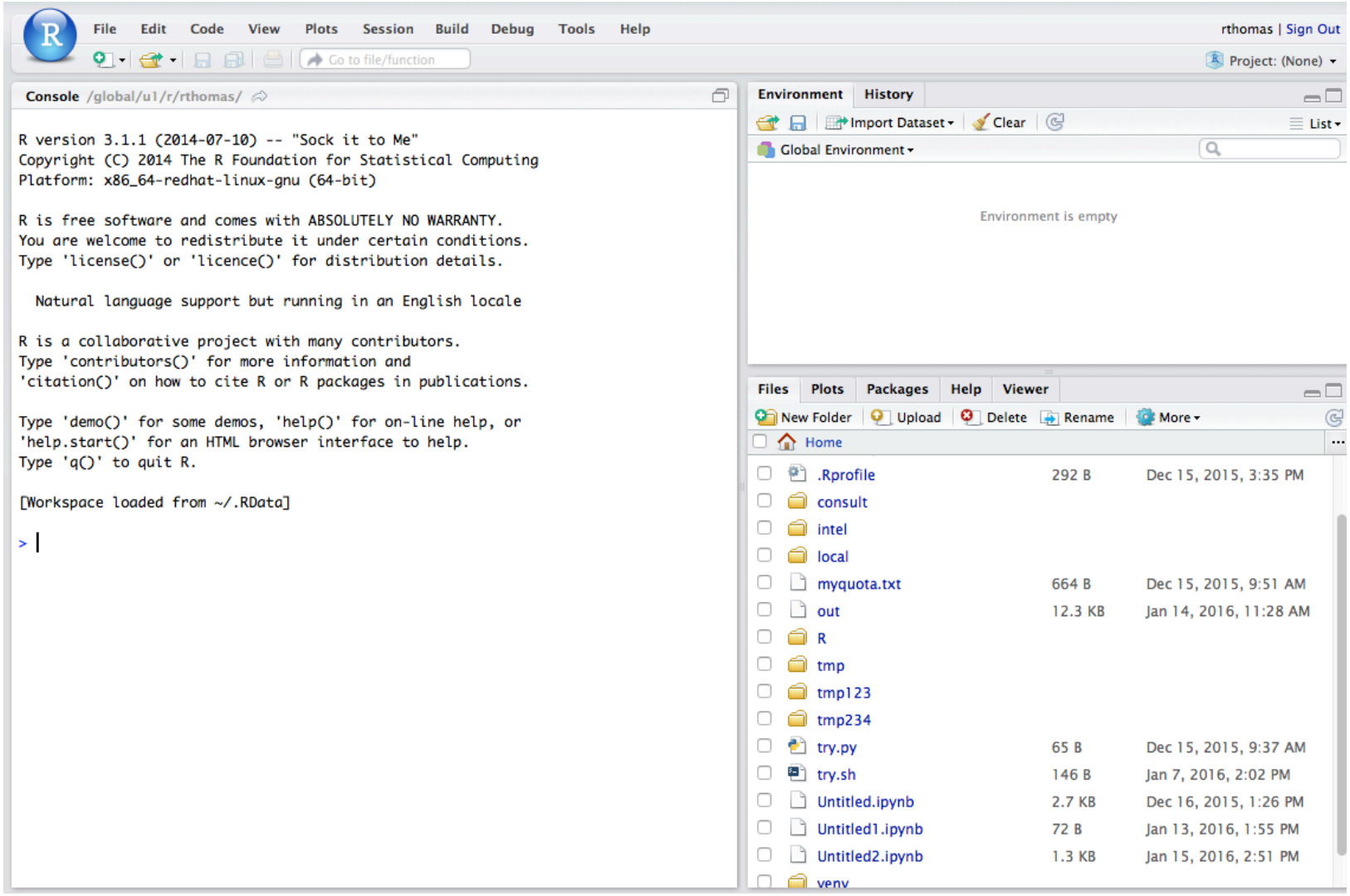
rstudio.nersc.gov

Just your usual NERSC username & password.

IPython


jupyter
Control Panel
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Text File
Folder
Terminal
Notebooks
Python 2
Python 3
consult
intel
local
tmp
tmp123
tmp234
venv
work
Untitled.ipynb
Untitled1.ipynb
myquota.txt
out
try.py
try.sh
Click to launch notebook...
jupyter Untitled2 Last Checkpoint: 3 minutes ago (unsaved changes)
Control Panel
Logout
File Edit View Insert Cell Kernel Help
Python 2
Code
Cell Toolbar: None
In [1]: import glob
glob.glob("*")
Out[1]: ['Untitled.ipynb',
'tmp',
'venv',
'intel',
'Untitled2.ipynb',
'Untitled1.ipynb',
'consult',
'myquota.txt',
'try.py',
'tmp234',
'out',
'local',
'work',
'try.sh',
'tmp123']
In [ ]:
Notebook can see:
$HOME
/project
/global/project{a,b}
/global/dna
Not $SCRATCH

RStudio

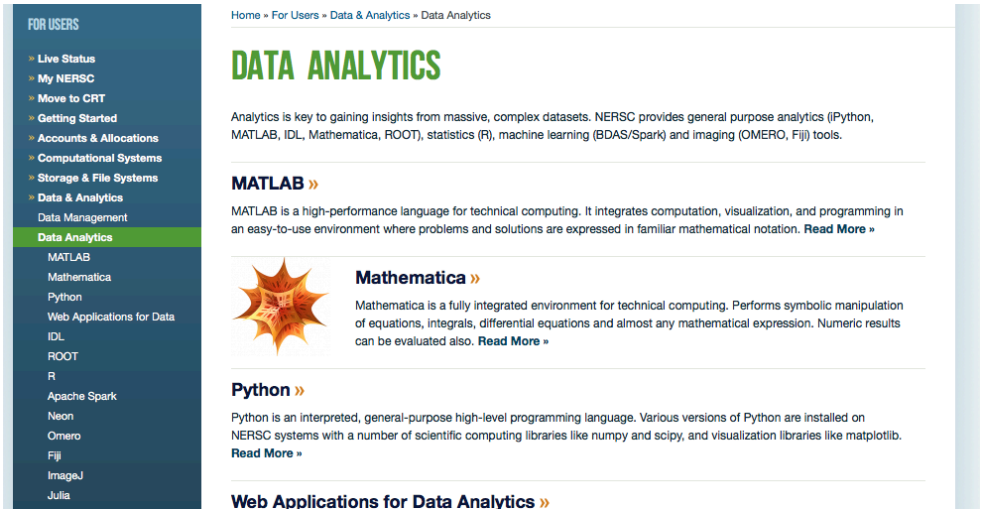

File Edit Code View Plots Session Build Debug Tools Help
rthomas | Sign Out
Project: (None)
Console /global/u1/r/rthomas/
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Environment History
Import Dataset
Clear
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder
Upload
Delete
Rename
More
Home
.Rprofile 292 B Dec 15, 2015, 3:35 PM
consult
intel
local
myquota.txt 664 B Dec 15, 2015, 9:51 AM
out 12.3 KB Jan 14, 2016, 11:28 AM
R
tmp
tmp123
tmp234
try.py 65 B Dec 15, 2015, 9:37 AM
try.sh 146 B Jan 7, 2016, 2:02 PM
Untitled.ipynb 2.7 KB Dec 16, 2015, 1:26 PM
Untitled1.ipynb 72 B Jan 13, 2016, 1:55 PM
Untitled2.ipynb 1.3 KB Jan 15, 2016, 2:51 PM
venv

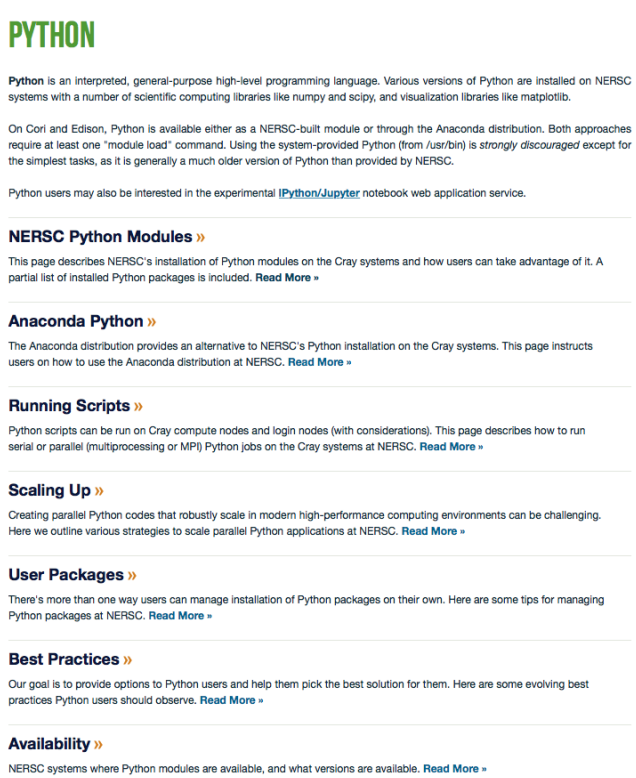
Getting Help with Data Analytics at NERSC

Ways to Get Help

- Ask us anything! Tell us anything! Suggest anything!
  - consult@nersc.gov
- Documentation:
  - http://www.nersc.gov/users/data-analytics/data-analytics/
  - http://www.nersc.gov/users/data-analytics/data-analytics/python/

  (Just reorganized and updated!)





Developments in Data Analytics at NERSC

Recent and Near-Term Developments

- **Jupyter on Cori: (mid-to-late 2016).**
  - Jupyter notebooks using Cori compute nodes.
    - Access to Cori $SCRATCH.
    - Use notebooks to launch large analytics workflows.
    - NERSC+: UC Berkeley, Cray, HPC community.
- **Consolidating Python package management:**
  - Update schedule tied with system upgrades.
  - Looking for user feedback and prioritization.
    - Ability to manage your own software stack?
    - Anaconda vs NERSC-built vs Intel distribution?
- **Julia, Spark on Cori:**
  - Available experimentally, please kick the tires!

Conclusion

- **Data Analytics**: The key to unlocking insight from massive and complex data sets.
- NERSC supports a variety of general-purpose analytics tools and services.
- What this talk covered:
  - Data analytics tools available on the Cray machines.
  - Other analytics services enabled through the web.
  - How to get help with data analytics at NERSC.
  - What's new and coming up in data analytics tools.

National Energy Research Scientific Computing Center