Using RStudio and Jupyter on the Cluster

Getting started with the High-Performance Compute Cluster

*Created by Kyle Monahan, April 2018*

*Last edited: 5-23-2018*

*Status: PUBLIC | For more information on statistical software support at Tufts, see* [*go.tufts.edu/stats/*](http://go.tufts.edu/stats/)

Contents

[1. Introduction 1](#_Toc511667065)

[2. Obtaining an account on the cluster 2](#_Toc511667066)

[3. Accessing the cluster 2](#_Toc511667067)

[3.1. Through the browser 2](#_Toc511667068)

[3.1.1. Files 3](#_Toc511667069)

[3.1.2. Jobs 3](#_Toc511667070)

[3.1.3. Clusters 3](#_Toc511667071)

[3.1.4. Interactive Apps: 4](#_Toc511667072)

[Jupyter and Jupyter Lab: 5](#_Toc511667073)

[RStudio: 5](#_Toc511667074)

[3.1.5. Miscellaneous: 6](#_Toc511667075)

[3.2. On Mac/Linux 7](#_Toc511667076)

[3.2.1. Using ssh 7](#_Toc511667077)

[3.2.2. Running x11 RStudio or Jupyter via ssh. 7](#_Toc511667078)

[3.2.3. Requesting an interactive node on the cluster 7](#_Toc511667079)

[Running RStudio from shell on Tufts HPC 8](#_Toc511667080)

[4.0 Uploading files to and from the cluster 9](#_Toc511667081)

[5.0 Learning more about the cluster 10](#_Toc511667082)

**Skills Covered in this Tutorial Include:**

* Accessing the cluster using the OnDemand interface,
* Loading RStudio and Jupyter notebooks on the cluster through command line
* Loading RStudio and Jupyter notebooks on the cluster through graphical interfaces

# Introduction

This tutorial is designed to get users started with using the RStudio integrated development environment (IDE) for students, faculty or staff at Tufts University.

RStudio is used to assist in writing code in the statistical programming language R. RStudio combines a powerful code/script editor for R code, visual tools for plotting and for viewing R objects and code history, a package manager for adding functionality, and a code debugger.

There are many methods to use RStudio software, including installing the software locally to your PC ([Rstudio.com](https://www.rstudio.com/products/RStudio/)), using a Tufts-owned computer such as those available in the Data Lab ([datalab.tufts.edu](http://datalab.tufts.edu/)), using the Tufts Virtual Desktop interface ([vdi.it.tufts.edu](http://vdi.it.tufts.edu/)), and finally using the Tufts High Performance Compute (HPC) cluster, herein referred to as the cluster. In this tutorial, we focus on the last method – assessing RStudio using the cluster.

This tutorial is suitable for those who have not worked with R/RStudio before. This tutorial may take 1 to 2 hours to complete.

# Obtaining an account on the cluster

To obtain an account on the cluster, you can follow the instructions below (adapted from [this page](https://it.tufts.edu/clusacct)):

To request a Research Computing Cluster account:

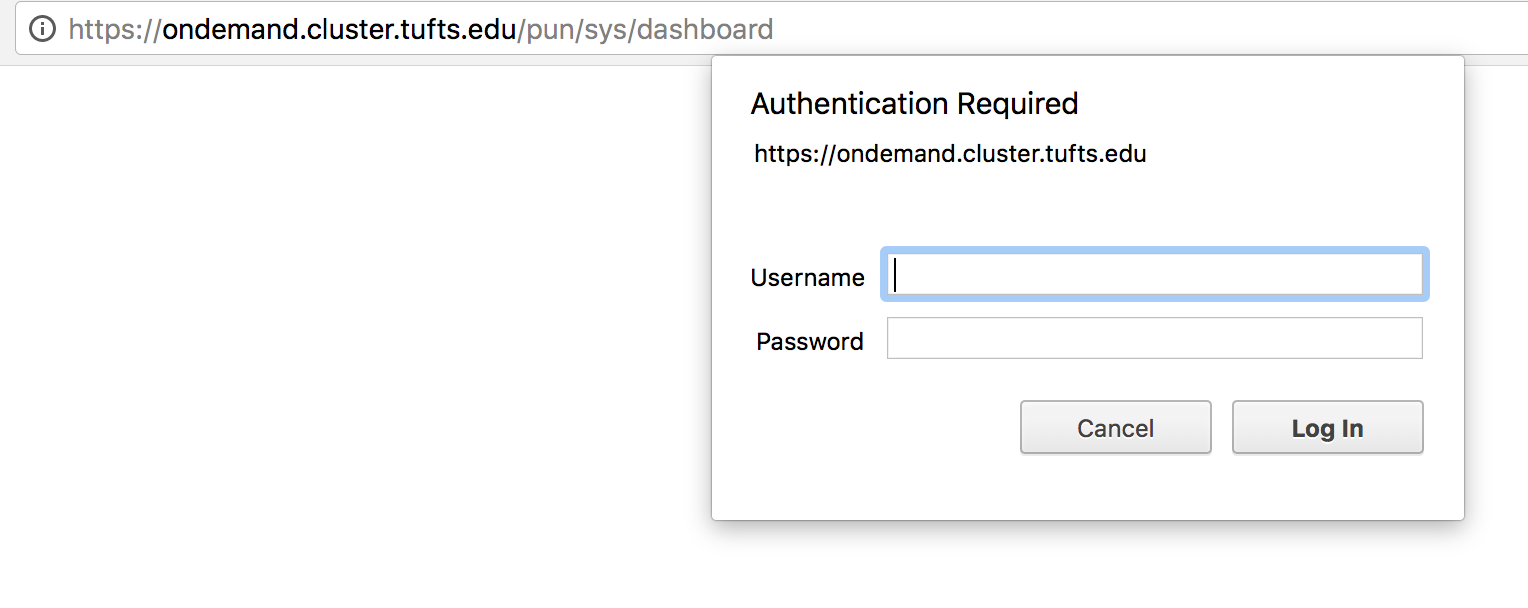
1. Browse to <http://research.uit.tufts.edu/> and select the Research Computing Cluster link.
2. Log in with your Tufts Username and Tufts Password.
3. The form will auto-fill as much of your user information as possible. Double-check to make sure it’s correct, selecting the down arrow to adjust the affiliation information if it’s incorrect.
4. Remove the example text in the Usage Information box and briefly describe your planned use of the Cluster.
5. In the Type of Account field, select the Cluster.
6. When finished, click, Submit Request.

# Accessing the cluster

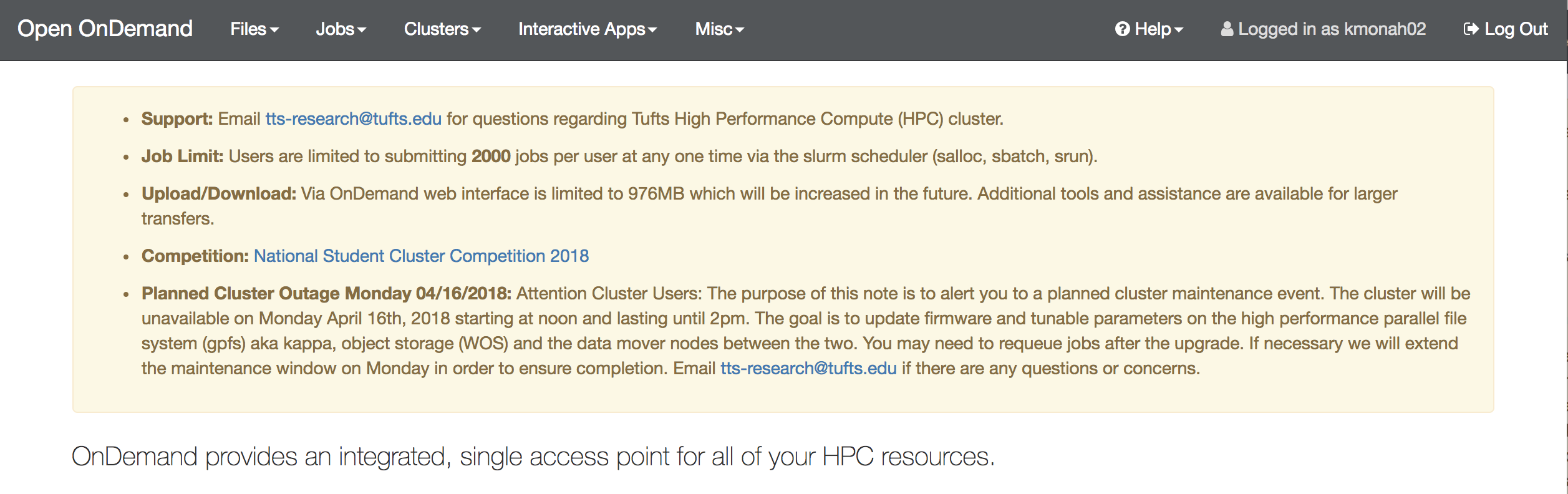
## Through the browser

The first way to access the cluster is through the Internet browser, such as Chrome, Safari, Edge (Internet Explorer), or Mozilla Firefox. After applying for an account, you can access the cluster through a browser by:

1. Navigating to ondemand.cluster.tufts.edu, or clicking the link here. In the Tufts Data Lab, just search for the program "Tufts HPC Cluster" at the start search bar.
2. A popup with the title “Authentication Required” will appear. Log in with your Tufts UTLN and password, and then click Log In.



1. You will then see the following menu:



### Files

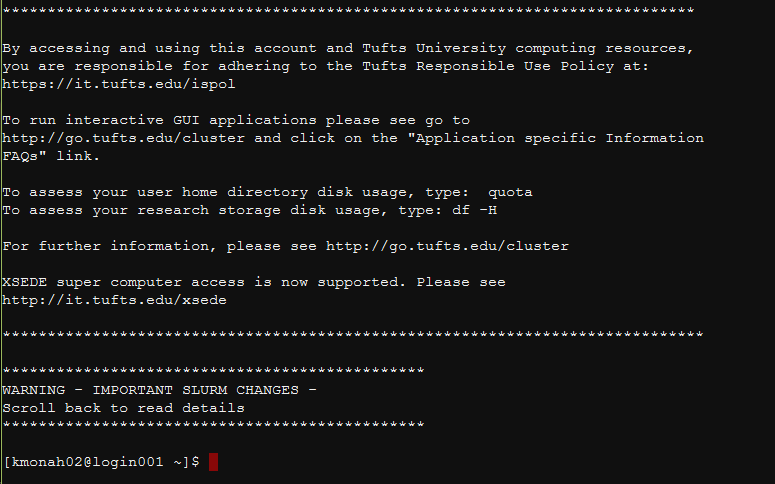
* + 1. This allows you to upload and download personal files from and to the cluster. The location of these files on the cluster is defaulted to /cluster/home/$USER/ where $USER is your username (Tufts UTLN). This is also the place where you start when you login to the cluster, your home directory.
       1. It’s important to note that there are other options for accessing data on the cluster. You can apply for research storage space by clicking the link below. [https://research.uit.tufts.edu/storage-request/html/index.php/auth/logi](https://research.uit.tufts.edu/storage-request/html/index.php/auth/login)n

### Jobs

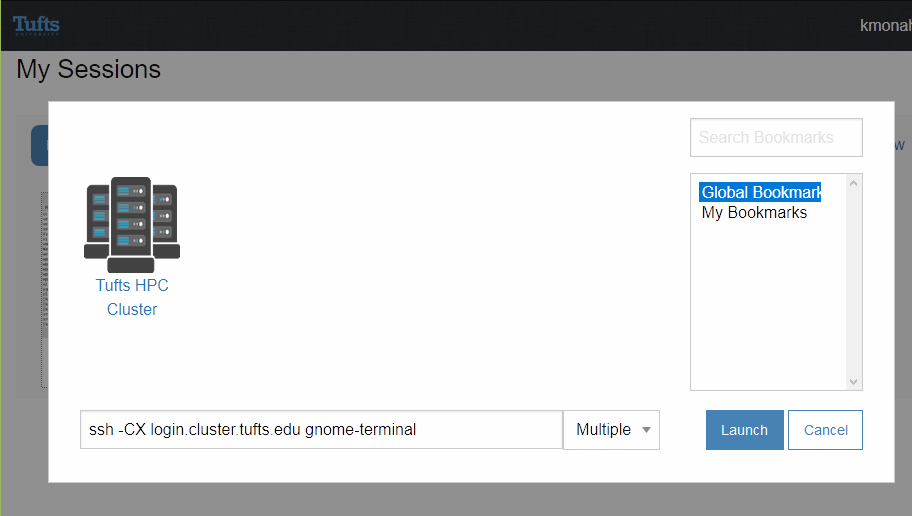
* + 1. When you submit a command to a clustered computer, you are likely not the only user submitting commands. To deal with the parallel nature of commands, the scheduler organizes the order in which it will process users' commands. If you click on *Jobs > Active Jobs,* you can see any jobs or commands you have sent, and their relative order in the queue of jobs.

### Clusters

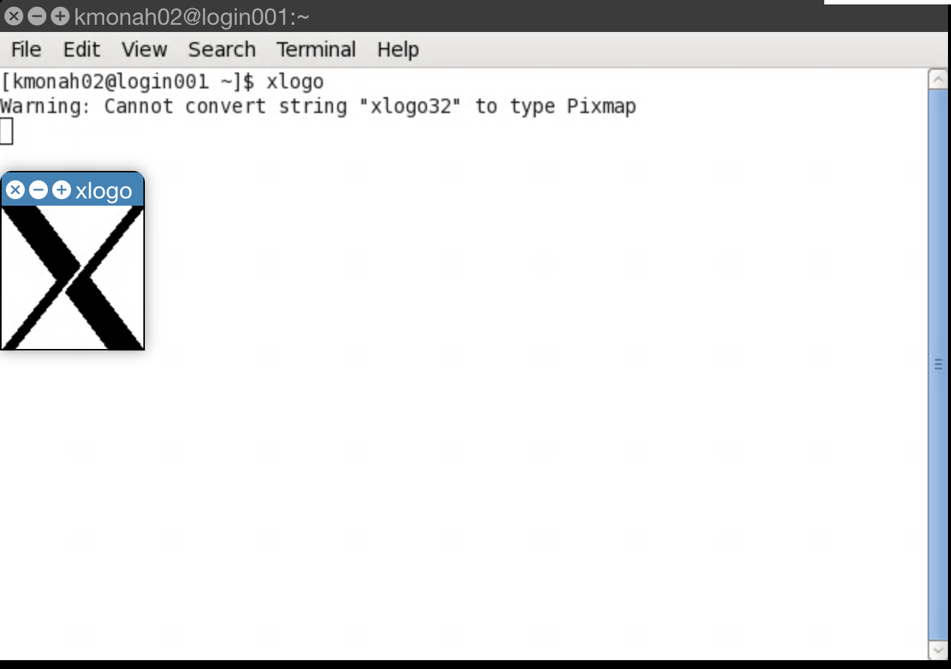
* + 1. **Tufts HPC Shell Access**: This will allow you to access the shell, or a text-entry command line interface, to the cluster. If you click on this button, a new tab will open and you will be dropped into a command interface on the cluster, able to run commands. For more information on what commands you can run, see here: <https://wikis.uit.tufts.edu/confluence/pages/viewpage.action?pageId=81760344>



* + 1. **Tufts HPC FastX Shell Access:** This is similar to the HPC shell, but adds additional FastX capacity. FastX refers to X11 graphics support. X11, or the X-Windows system is a Unix graphical user interface. For more information, see here: <https://en.wikipedia.org/wiki/X_Window_System>
       1. To use this, click on FastX Shell Access. Login with your Tufts UTLN (username) and password, as normal. Then click on *Launch Session > Click the Tufts HPC Cluster icon > Click Launch*. After that, a cluster icon will appear in the FastX window. Click on it to open a shell with FastX support in a new window.

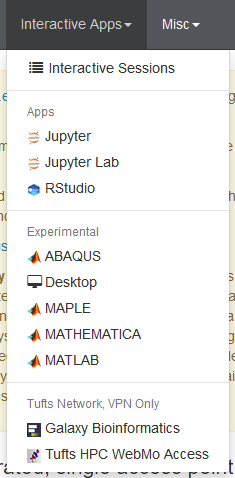


* + - 1. If you want to confirm that X11 graphics are working, you can type the following at the shell command window:
         1. *xlogo*
      2. If X11 graphics are working, you should see an X graphic appear, as in the screenshot below.

**

### Interactive Apps:

* + 1. The cluster has interactive apps, which allow for in-browser access to selected tools through their graphical interfaces. These include Jupyter, Jupyter Lab, RStudio, Maple, MATLAB, Galaxy Bioinformatics, and WebMo.



#### Jupyter and Jupyter Lab:

* + 1. Jupyter is the successor to iPython notebooks, and allows users to write R and Python on the High-Performance Compute Cluster. New users should use Jupyter Labs due to the easy to use interface.
       1. **Loading Jupyter Lab or Jupyter**: Click on Jupyter Lab/Jupyter under Interactive Apps. Input the number of hours you would like to be able to access the instance of Jupyter Lab through the cluster, and press *Launch*.
       2. **Home directory**: The default location to save your work is the users home directory (/cluster/home/$USER/ where $USER is your username (Tufts UTLN)).

#### RStudio:

* + 1. RStudio is an integrated development environment (IDE) for the R statistical programming language. This allows you to write R code, manipulate data, and view outputs using a graphical interface.
       1. **Loading RStudio**: Click on RStudio under Interactive Apps.
          1. Input the number of hours, number of nodes, and node type that you need.

**Number of hours**: This is how long you require the session of RStudio to be running. Your files can be saved during a session and available afterwards, but you must save your files before the end of the time. To prevent requesting more time than you need, we suggest you request 1-2 hours to start out.

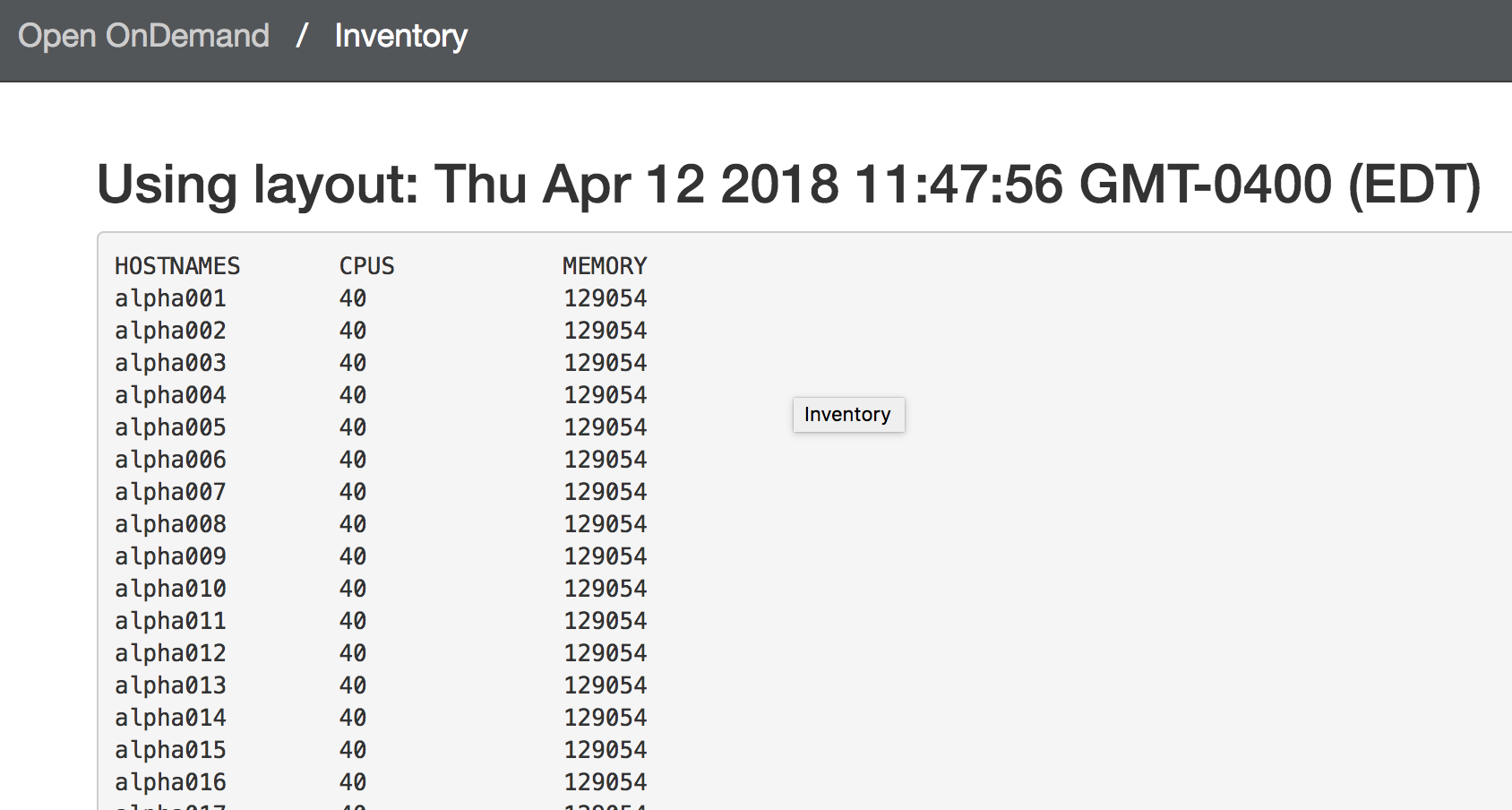
**Number of nodes**: Each node has access to anywhere from 24 – 48 compute cores. If you require a process that runs in parallel across two or more nodes, or require a high number of compute cores, then you can request more nodes. Remember that this may increase your wait time in the scheduling system!

**Node type**: If you are in need of a large memory node (lrgmem), you can request a *lrgmem* node, which contains 48 Haswell cores and 1TB of available RAM. However, if you do not need to work with large memory processes, you do not need to use this node, and can otherwise select all.

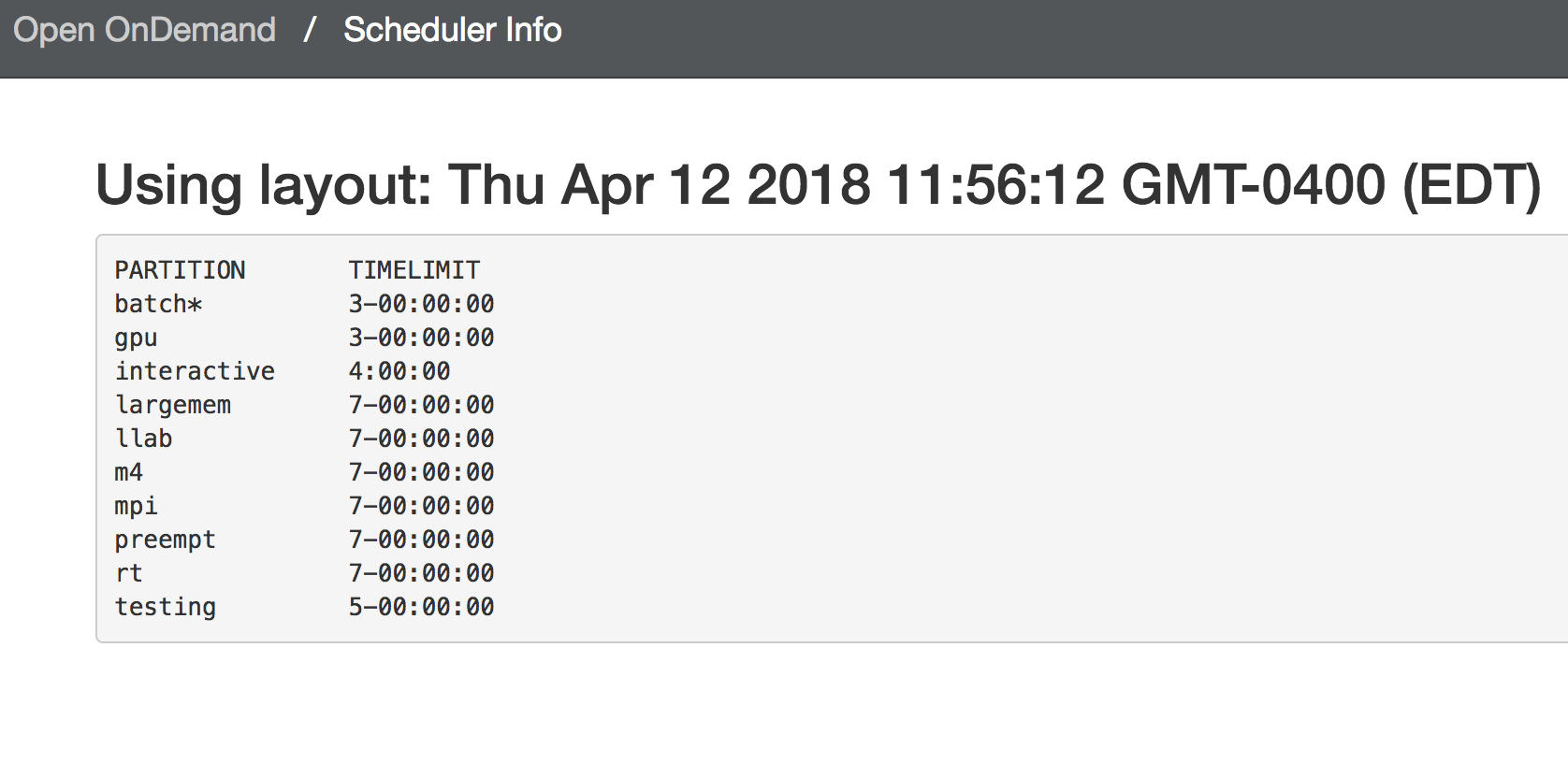
* + 1. **Experimental Apps:** Other experimental apps exist, but are beyond the scope of this tutorial.

### Miscellaneous:

* + 1. **Inventory**: This allows users to view all nodes, their CPU allocation, and the amount of memory available. This corresponds with the



* + 1. **Quota Report**: For storage on the cluster, each user has a quota or maximum storage provided for each group the user is a member of. group *student*.
    2. **Scheduler Info:** This gives you an idea of each partition of nodes, and the time limit for each partition.



* + 1. Show Groups: If you wonder what groups you are a part of e.g. if you are a student, you are likely in group *student* - you can check that here.

## On Mac/Linux

### Using ssh

On both Mac and Linux operating systems, a utility called Secure Shell (*SSH*) exists. This is a way to encrypt the connection between a local computer (such as a Mac or Linux PC) and a remote computer (such as the Tufts HPC cluster). You’ll need a Tufts HPC account first, (Section 2.0 titled “Obtaining an Account on the Cluster”).

To connect to the Tufts HPC, you can issue the following command from Terminal:

*ssh -YC $UTLN@login.cluster.tufts.edu*

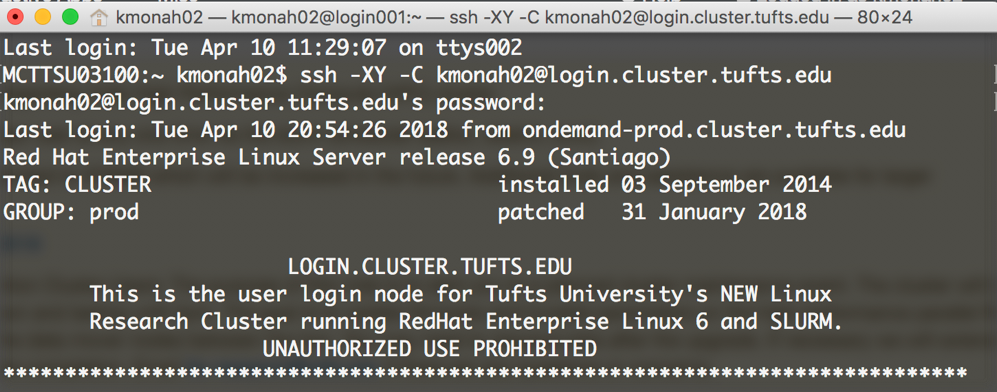
*Where:*

*ssh the function you are using, secure shell – see the man(ual) [page here](http://linuxcommand.org/lc3_man_pages/ssh1.html)*

*-YC these are called flags, Y enables X11 graphical forwarding, and C requests compression*

*$UTLN This is your tufts username, e.g. kmonah02*

You can then input your Tufts password. Note that no characters will appear, to protect your security. Press enter and you will have a shell on the cluster, screenshot below.



### Running x11 RStudio or Jupyter via ssh.

After logging into cluster using *ssh* (Section 3.2.1) or a similar shell client, one can run both RStudio and Jupyter from *ssh*. This is the same process as running any command on the shell, but this requires us to do the following:

1. Pass the *ssh* function the -Y flag when logging in, to enable x11 forwarding.
2. You must install an x11 client such as xQuartz for Mac <https://www.xquartz.org/>
3. Finally, you must request an interactive session on the cluster.

### Requesting an interactive node on the cluster

When you work via a shell (such as *ssh*) on the cluster, you **must** request an interactive node before running an analysis, including using RStudio or Jupyter. This is because when you start on the cluster, you are actually on the login node, as below, [$UTLN@login001 ~]:



We need to request an interactive node on the cluster, which can support our plan to run analysis, commonly called jobs, in an iterative and interactive fashion. To do this, we can run a line of code. This code asks the scheduler (called SLURM, more information here: <https://wikis.uit.tufts.edu/confluence/display/TuftsUITResearchComputing/Slurm)>.

**Table 1**. A list of common SLURM commands

|  |  |
| --- | --- |
| **Task** | **SLURM**  **Command** |
| kill a job | scancel |
| what are my running job(s) | squeue |
| submit a job | sbatch |
| what queues/partitions | squeue |
| how to obtain resources | salloc |
| node listing | sinfo |
| interactive session or application | srun |
| controlling jobs | scontrol |
| usage accounting | sreport, sacct, sstat |
| other submissions | srun |
| copy files to allocated local disk | sbcast |

In this case, we want to create an interactive session, so we will use the *srun* command. The command is:

*srun -p interactive --pty --x11=first --mem=4000 bash*

*Where:*

*srun the function you are using, SLURM command for interaction sessions*

*-p this is the type of node you are requesting, called a partition – a*

*--pty pseudo-terminal request – asks for a shell (command) line on the interactive node*

*--x11=first this asks for an x11 graphical connection on that shell (command) line, there are [other options](http://portal.hpcwales.co.uk/wordpress/index.php/index/slurm/interactive-use-job-arrays/x11-gui-forwarding/)*

*--mem=4000 this requests 4000 MB or about 4GB of RAM on the node*

*bash this is the type of shell (command) line we are requesting, in this case it is a bash shell*

**Table 2**. A list of all Tufts HPC partitions as of 4-13-2018.

|  |  |
| --- | --- |
| **PARTITION** | **TIMELIMIT** |
| batch | 3-00:00:00 |
| gpu | 3-00:00:00 |
| interactive | 4:00:00 |
| largemem | 7-00:00:00 |
| llab | 7-00:00:00 |
| m4 | 7-00:00:00 |
| mpi | 7-00:00:00 |
| preempt | 7-00:00:00 |
| rt | 7-00:00:00 |
| testing | 5-00:00:00 |

#### 3.2.4 Running RStudio from shell on Tufts HPC

Now that we are on an interactive node (Section 3.2.3) on the cluster, we can run RStudio.

First, we purge all modules: load R and RStudio and run Rstudio

*module purge*

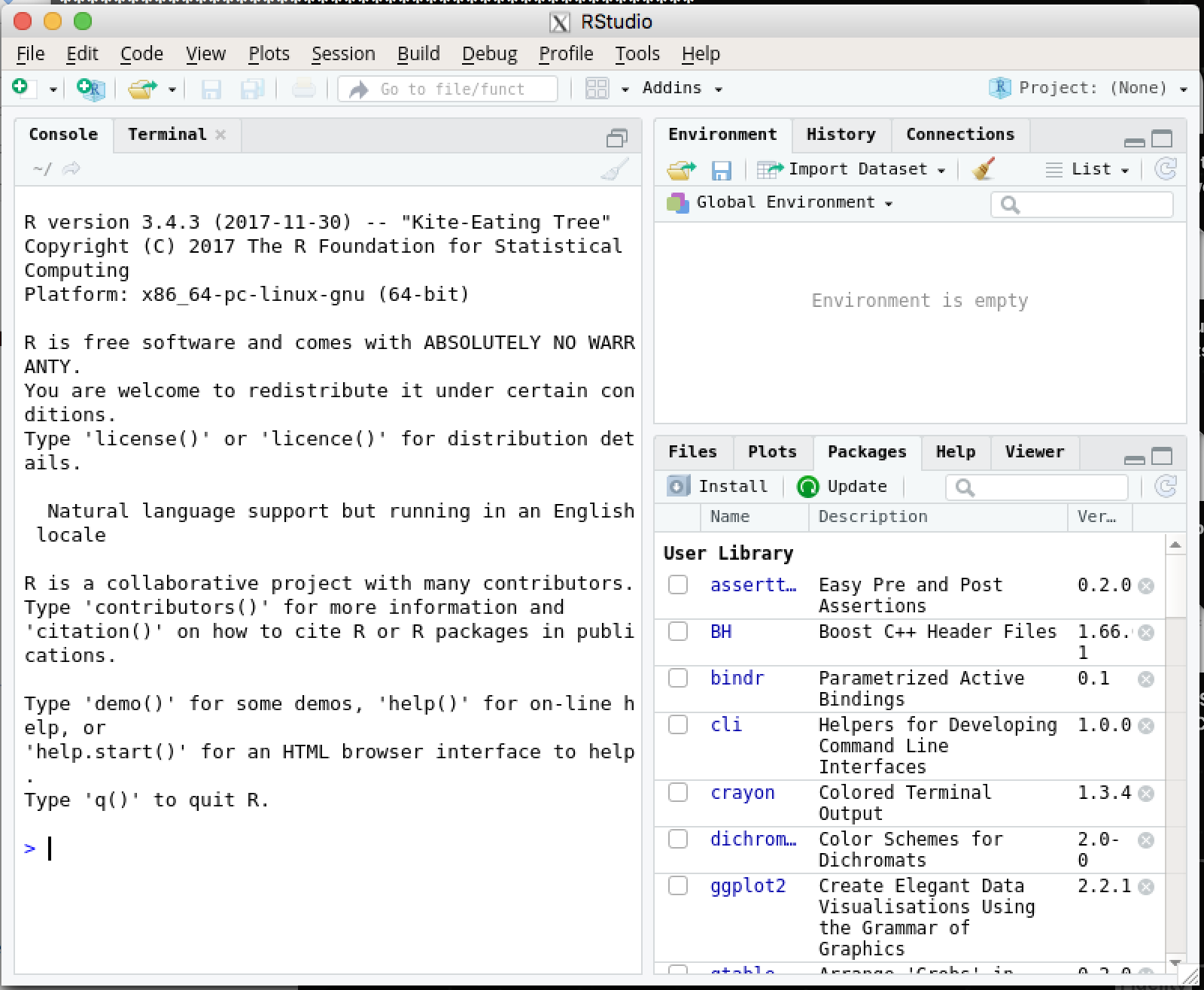
Then we load R, RStudio and run RStudio:

*module load R*

*module load rstudio/1.1.383*

*rstudio*

This will produce a pop-up window, similar to the image below. This allows you to use RStudio just like you would on your local machine.



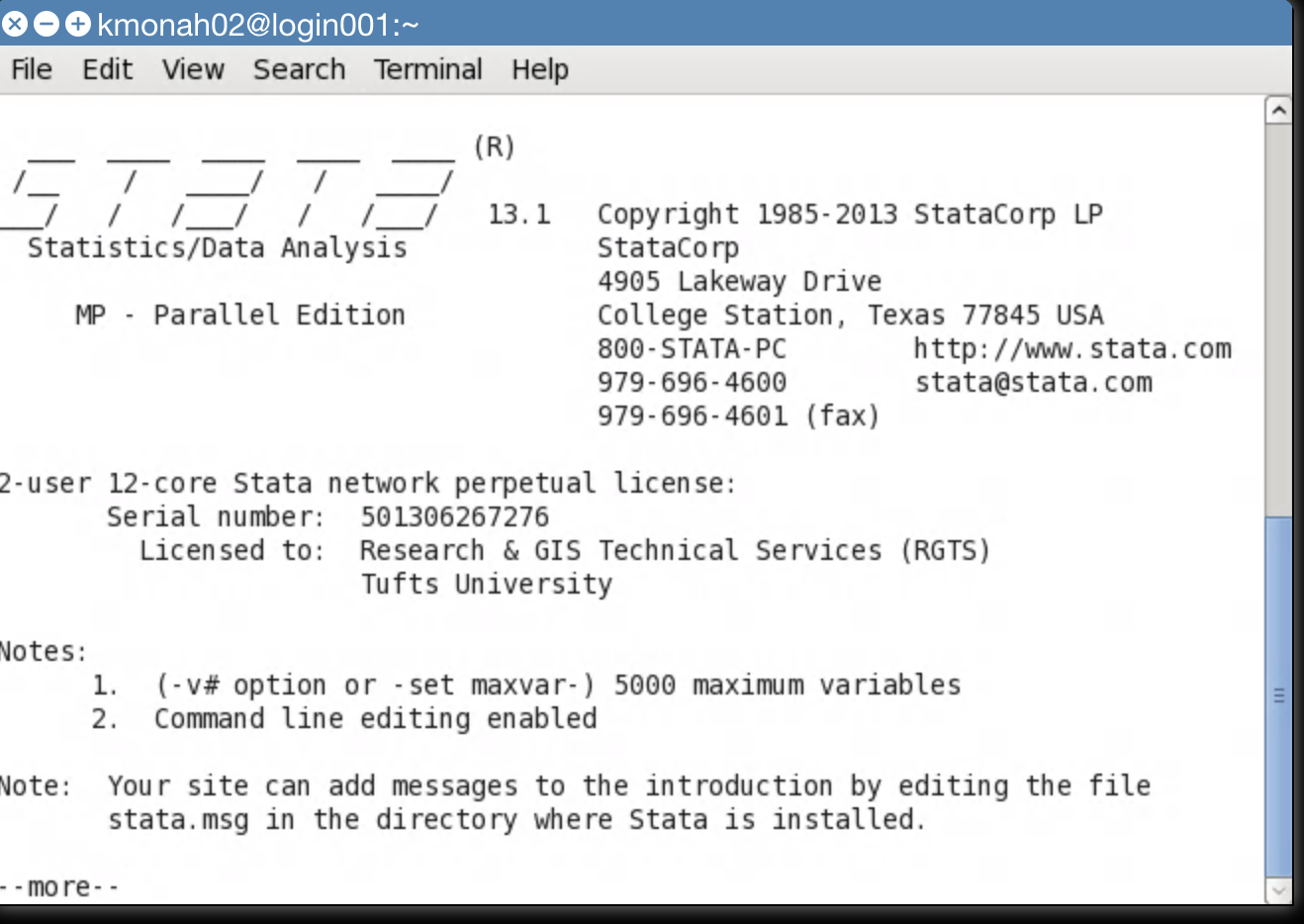
#### 3.2.5 Running Stata from shell on Tufts HPC

You can also run other programs, including Stata, in the same manner. For example, to load the high performance Stata MP, type the following:

*module purge*

*module load stata/13*

*srun --pty -p interactive stata-mp #Note you can modify this command using the srun syntax above in 3.2.3.*



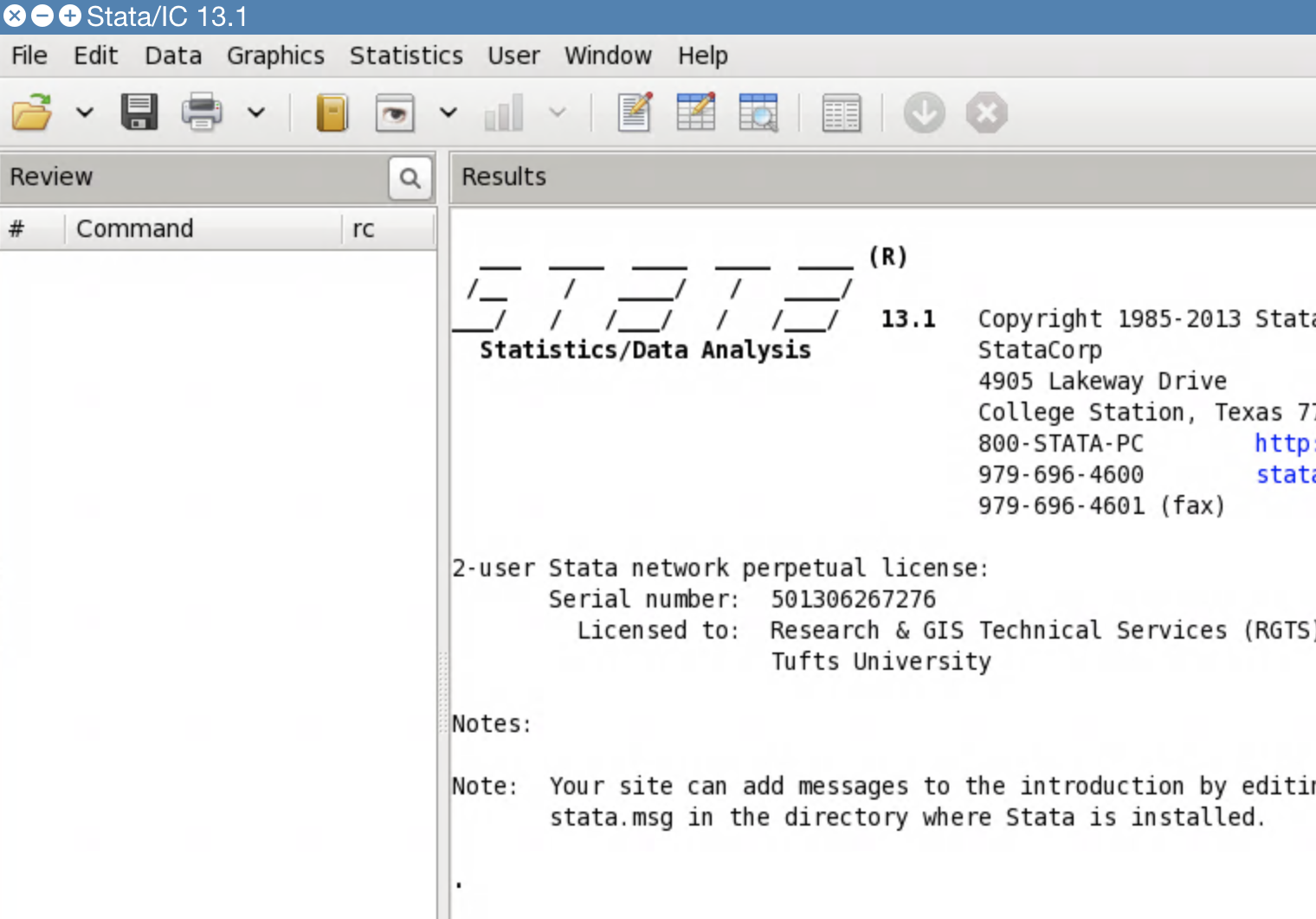
If you prefer the lower performance, but graphical Stata IC, you can type:

*module purge*

*module load stata*

*xstata*

And a graphical window will appear, as below.



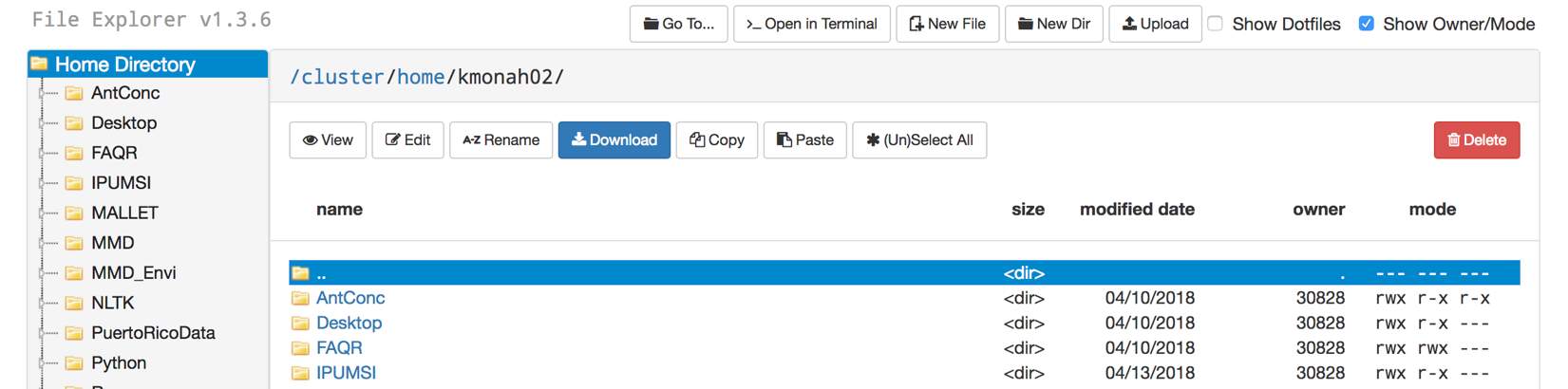
# Uploading files to and from the cluster

To upload files to the cluster, you can do the following:

1) Navigate to go.tufts.edu/OnDemand and login with your Tufts username and password when prompted.

2) Then, navigate to the *Files* tab in the upper left. Click “*Home directory*” under the Files tab.

3) By clicking *Download*, you can download files. By clicking *Upload* at the upper right, you can upload files.



# 5.0 Learning more about the cluster

To learn more about the cluster, please navigate to the following links:

* Tufts HPC Cluster: <https://wikis.uit.tufts.edu/confluence/display/TuftsUITResearchComputing/Home>
* New HPC Users: <https://wikis.uit.tufts.edu/confluence/display/TuftsUITResearchComputing/New+HPC+users>
* Livermore Computing Center Training: <https://hpc.llnl.gov/training/tutorials>

If you have any questions, contact Kyle Monahan at kyle.monahan@tufts.edu.