
# JUPYTER ASCENDING:
## A PRACTICAL HAND GUIDE TO GALACTIC SCALE, REPRODUCIBLE DATA SCIENCE

John Fonner, PhD
University of Texas at Austin

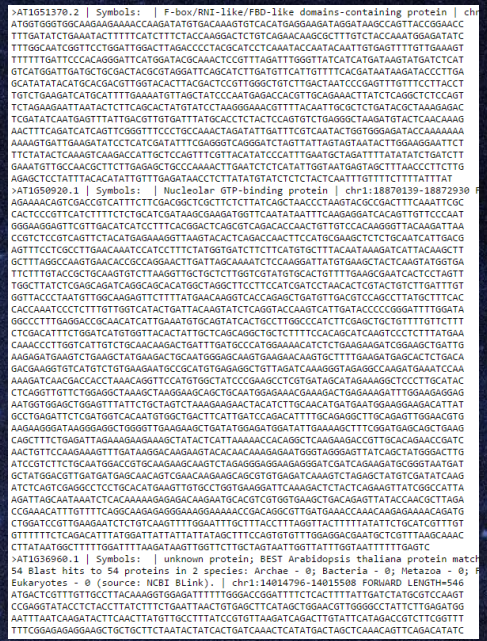
April 5th, 2016

- Photos, Tweets, and hate mail all welcome!
- Slides: tinyurl.com/FonnerSEA2016
- Email: jfonner@tacc.utexas.edu
- Twitter: @johnfonner

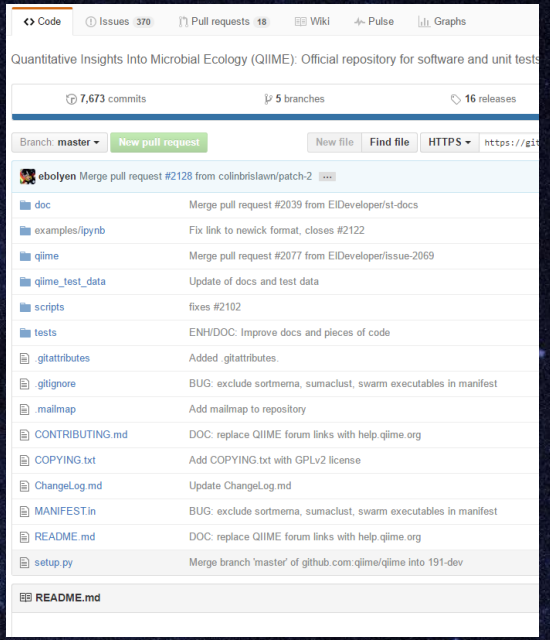
1. Formulate a theory
2. Gather data
3. Learn about data storage
4. Learn about data movement protocols
5. Lose data
6. Check out of rehab
7. Learn about backup and replication
8. Gather data
9. Learn about versioning
10. Start preliminary analysis
11. Buy a newer laptop
12. Buy more memory
13. Buy a desktop with more memory
14. Buy a bigger monitor & GPUs “for work”
15. Google “250GB Excel Spreadsheet”
16. Learn about batch processing
17. Learn about batch schedulers
18. Learn about patience.
19. Learn more about data storage
20. Learn about distributed systems.
21. Go back through notes to remember the science question.
22. Learn R & Python
23. Learn linux admin
24. Finish preliminary analysis.
25. Grow a ponytail
26. Write a paper.
27. Learn about data publishing
28. Learn about reproducibility
29. Plot the death of your advisor/dept. head
30. Apply for grants & research allocations on public systems
31. Wait to apply next time
32. Finish analyzing data
33. Reformulate your theory
34. Goto 1

# SCIENCE AS A SECOND THOUGHT


+



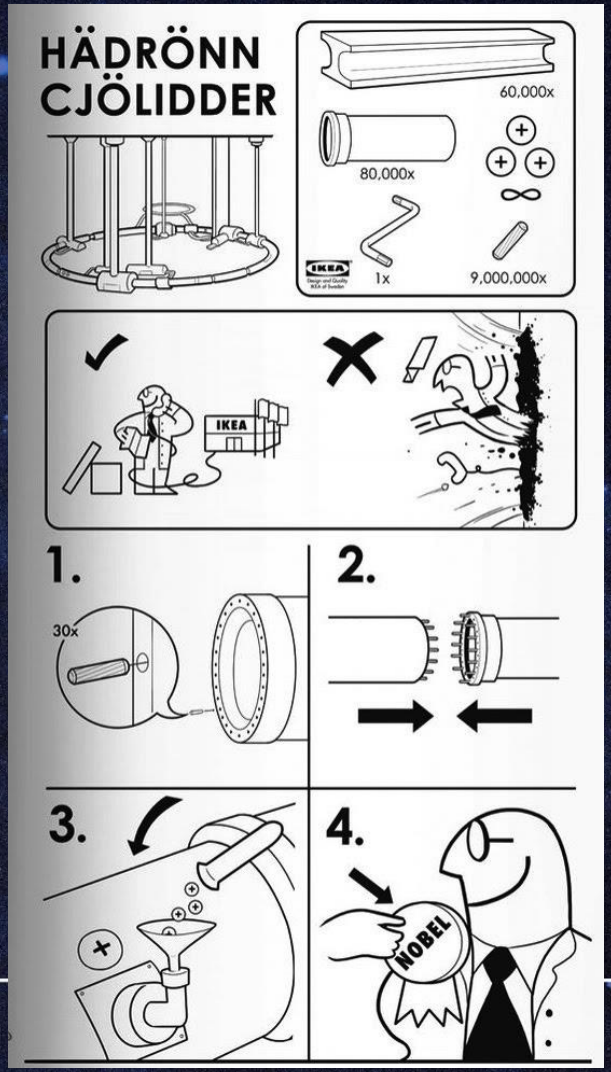

+



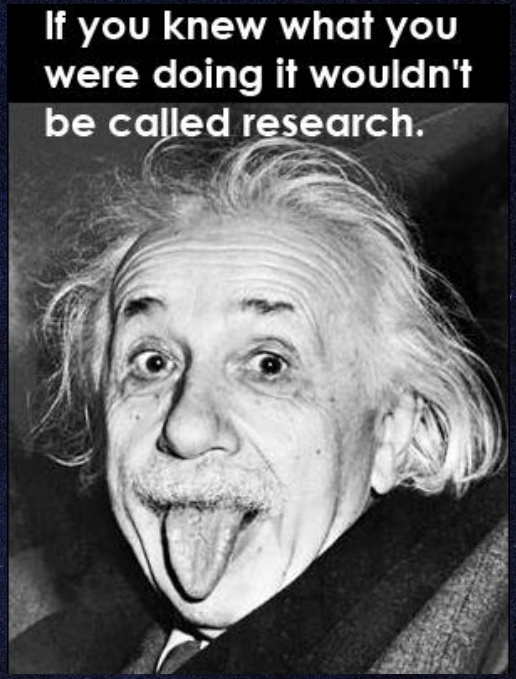

# SCIENTIFIC REPRODUCIBILITY




# SOME ASSEMBLY REQUIRED…


HÄDRÖNN
CJÖLIDDER
60,000x
80,000x
1x
9,000,000x
IKEA
1.
30x
2.
3.
4.
NOBEL

- Research is hard
- Coding is hard
- Research code is
  - ~~well designed~~,
  - ~~documented~~,
  - ~~leverages design patterns~~,
  - ~~highly reusable~~,
  - ~~portable~~,
  - and usually open source.




# SCIENTISTS, WITH FEW EXCEPTIONS, ARE ***NOT TRAINED PROGRAMMERS***

- For scientific reproducibility, the impact of your work will be more about accessibility than capability
  - Domain grad students, not sys admins, are the early adopters
  - Where can we focus effort to create community around capability?

# ACCESSIBILITY >= CAPABILITY

- What has changed the least about the computation you do over the last 10 years?

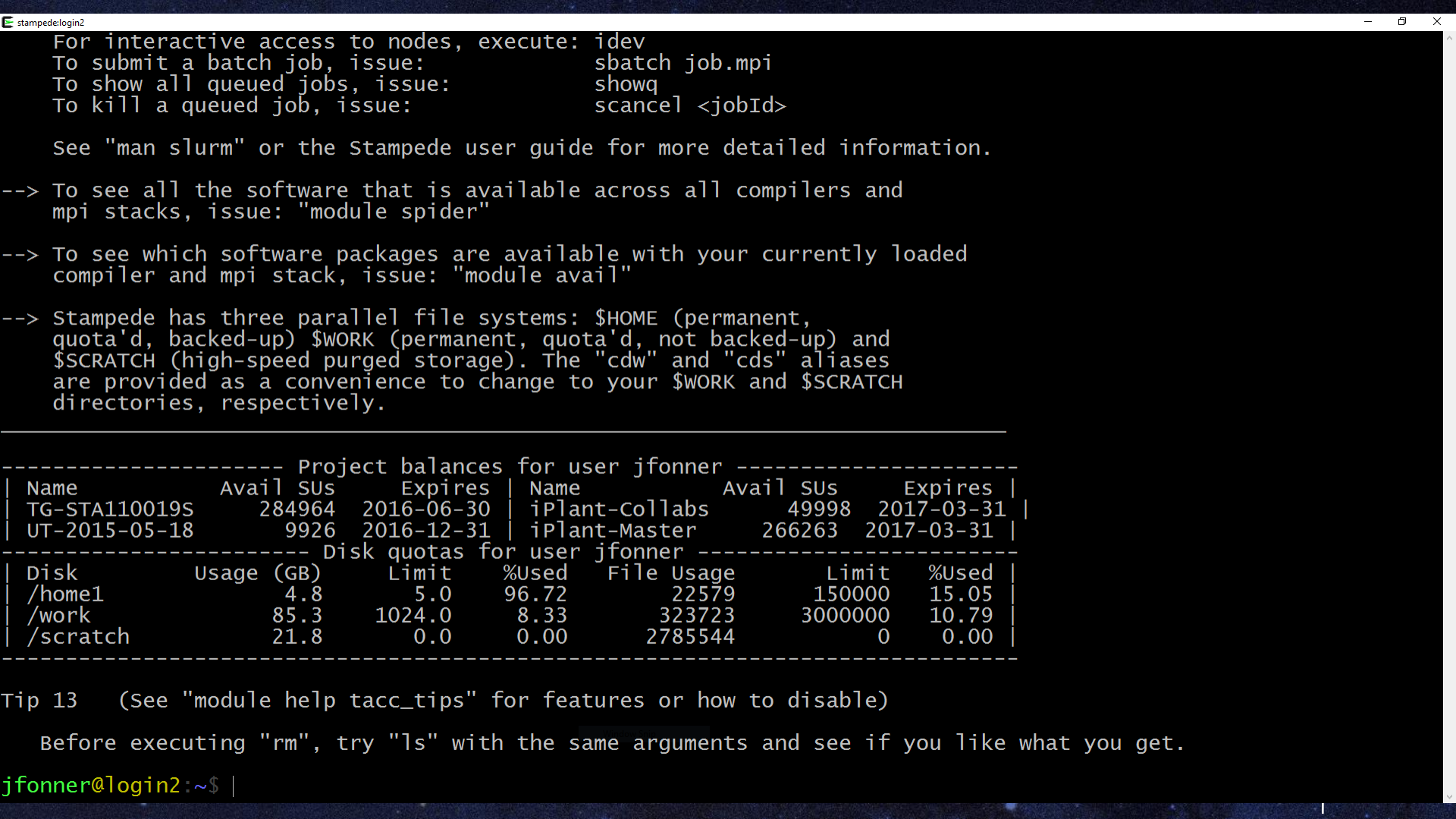
Memory/CPU/Disk

Operating System

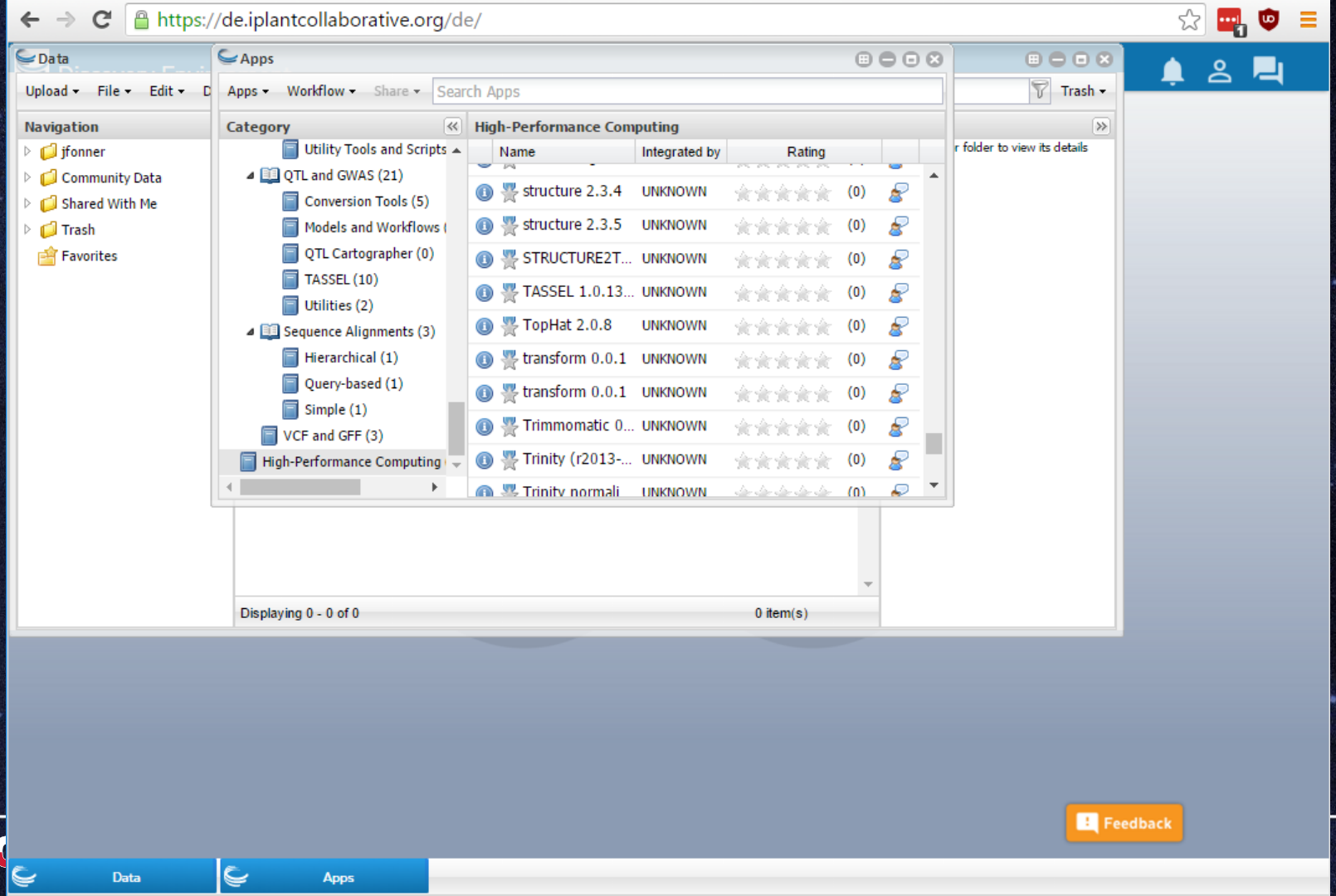
Applications

Interface

- What do we ask domain researchers to learn to use our tools and data?


stampede:login2
For interactive access to nodes, execute: idev
To submit a batch job, issue: sbatch job.mpi
To show all queued jobs, issue: showq
To kill a queued job, issue: scancel <jobId>
See "man slurm" or the Stampede user guide for more detailed information.
--> To see all the software that is available across all compilers and mpi stacks, issue: "module spider"
--> To see which software packages are available with your currently loaded compiler and mpi stack, issue: "module avail"
--> Stampede has three parallel file systems: $HOME (permanent, quota'd, backed-up) $WORK (permanent, quota'd, not backed-up) and $SCRATCH (high-speed purged storage). The "cdw" and "cds" aliases are provided as a convenience to change to your $WORK and $SCRATCH directories, respectively.
Project balances for user jfonner
Name Avail SUs Expires
TG-STA110019S 284964 2016-06-30
UT-2015-05-18 9926 2016-12-31
Name Avail SUs Expires
iPlant-Collabs 49998 2017-03-31
iPlant-Master 266263 2017-03-31
Disk quotas for user jfonner
Disk Usage (GB) Limit %Used File Usage Limit %Used
/home1 4.8 5.0 96.72 22579 150000 15.05
/work 85.3 1024.0 8.33 323723 3000000 10.79
/scratch 21.8 0.0 0.00 2785544 0 0.00
Tip 13 (See "module help tacc_tips" for features or how to disable)
Before executing "rm", try "ls" with the same arguments and see if you like what you get.
jfonner@login2:~$


ONE DOES NOT SIMPLY
DO DATA SCIENCE
imgflip.com


https://de.iplantcollaborative.org/de/
Data
Apps
Upload
File
Edit
Apps
Workflow
Share
Search Apps
Trash
Navigation
jfonner
Community Data
Shared With Me
Trash
Favorites
Category
Utility Tools and Scripts
QTL and GWAS (21)
Conversion Tools (5)
Models and Workflows
QTL Cartographer (0)
TASSEL (10)
Utilities (2)
Sequence Alignments (3)
Hierarchical (1)
Query-based (1)
Simple (1)
VCF and GFF (3)
High-Performance Computing
High-Performance Computing
Name
Integrated by
Rating
structure 2.3.4 UNKNOWN (0)
structure 2.3.5 UNKNOWN (0)
STRUCTURE2T... UNKNOWN (0)
TASSEL 1.0.13... UNKNOWN (0)
TopHat 2.0.8 UNKNOWN (0)
transform 0.0.1 UNKNOWN (0)
transform 0.0.1 UNKNOWN (0)
Trimmomatic 0... UNKNOWN (0)
Trinity (r2013-... UNKNOWN (0)
Trinity normali UNKNOWN (0)
folder to view its details
Displaying 0 - 0 of 0
0 item(s)
Feedback
Data
Apps

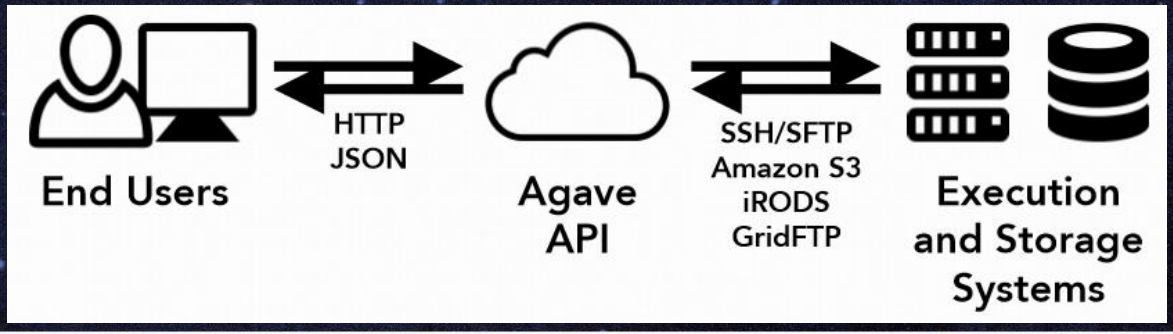
- Decoupling the technology "stack"

**"Reproducers"**

- **Web Browser**
- **GUIs**
- **Windows / Mac OS Support**
- **Sample Data and Sample Workflows**

**"Producers"**

- **Linux CLI**
- **Hadoop / GPFS / Lustre**
- **Clusters / Clouds / Containers**
- **Dockerfile / Makefile / Ansible**



- Categorize systems as either Storage or Execution
- Describe and support relevant protocols, directories, schedulers, and quotas
- Each system includes the credentials to log into the system (SSH Keys, X509, username/password)
- Register everything with a JSON document

http://agaveapi.co/documentation/tutorials/system-management-tutorial/

# BACKEND INFRASTRUCTURE: SYSTEMS

- An "App" is a versioned instance of a software package on a specific Execution System
- App assets are bundled into a directory and stored on a Storage System
- Apps can be private, shared with individual users, or made public
- Public apps are compressed, assigned a checksum, and stored in a protected space

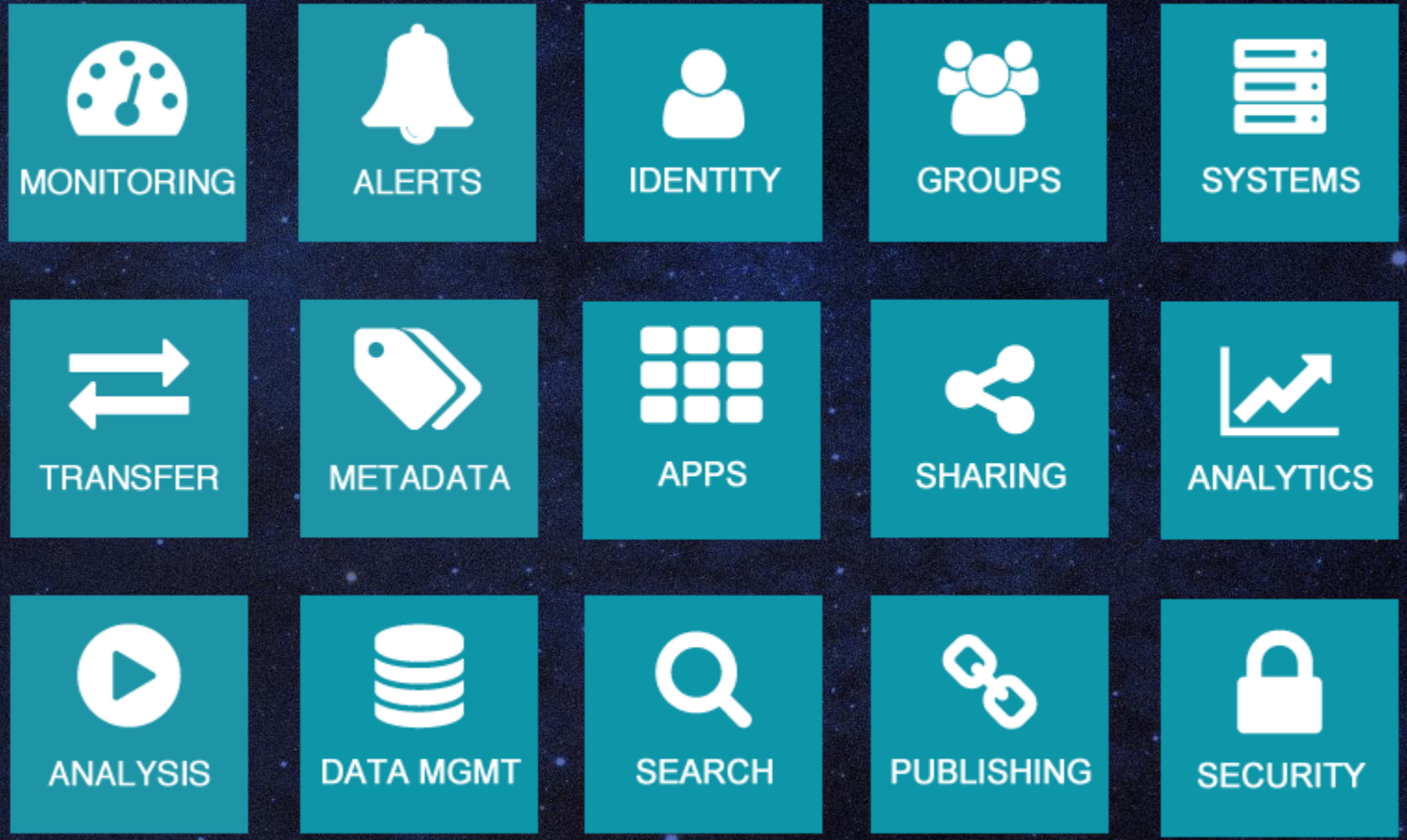
http://agaveapi.co/documentation/tutorials/app-management-tutorial/

# BACKEND INFRASTRUCTURE: APPS

- A "Job" is an execution of an App with a specific set of input files and parameters
- All jobs are given an ID, all inputs and parameters are preserved, output is also tracked
- Jobs can be shared with others

http://agaveapi.co/documentation/tutorials/job-management-tutorial/

# BACKEND INFRASTRUCTURE: JOBS

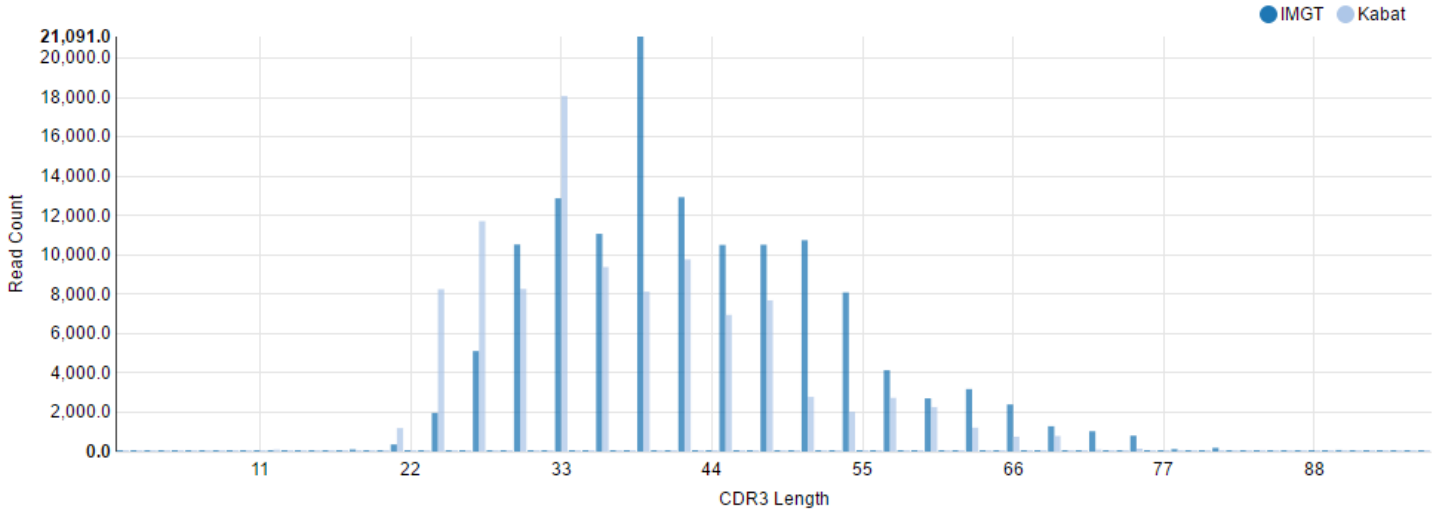

MONITORING
ALERTS
IDENTITY
GROUPS
SYSTEMS
TRANSFER
METADATA
APPS
SHARING
ANALYTICS
ANALYSIS
DATA MGMT
SEARCH
PUBLISHING
SECURITY

project 1
Testing Data_from Lindsay 2015.08.26
test 2015.07.27
demultiplex multibarcode
primers 2015.05.27
demultiplexing 2015.05.27
demultiplexing
SRP042205
Project Settings
Upload and Browse Project Data
Link .fasta/.qual Files
View Analyses and Results

| | | |
|---|---|---|
| SRR1297001.fastq | 143.39 MB | File Available in Project Data Area |

## Chart Files

| Name | Size | Chart |
|---|---|---|
| SRR1297001.fasta.igblast.out.cdr3_hist.tsv | 834 B | Hide Chart / Download Chart |



https://drugdiscovery.tacc.utexas.edu/#/

Drug Discovery @ TEXAS ADVANCED COMPUTING CENTER
*Powering Discoveries That Change The World*

THE UNIVERSITY OF TEXAS AT AUSTIN

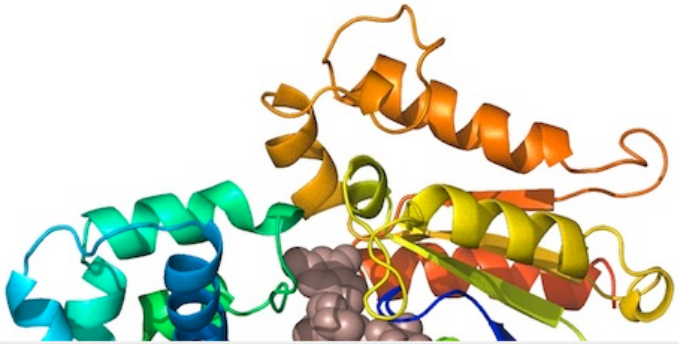
Home Data Jobs About FAQ

Username Password Login

# Welcome to the *new* Virtual Drug Discovery Portal!

This Portal provides a graphical interface for conducting a screen for identifying small molecules that bind to your target protein.

This new release is still in Beta.



https://www.designsafe-ci.org

DESIGNSAFE-CI
A NATURAL HAZARDS ENGINEERING COMMUNITY

Log in | Register

A CLOUD-BASED ENVIRONMENT FOR RESEARCH IN NATURAL HAZARDS ENGINEERING

NHERI Community | Research Workbench | NHERI Facilities | Learning Center | About | Contact

**NHERI COMMUNITY**

Relevant news, field-based opportunities, and user-guided discussions aimed at bringing the natural hazards engineering community together.

**RESEARCH WORKBENCH**

A comprehensive cloud-based research environment for experimental, theoretical, and computational engineering and science.

**NHERI FACILITIES**

Shared-use sites including Experimental Facilities, the Computational Modeling and Simulation Center, and the Network Coordination Office.

**LEARNING CENTER**

Training resources, site support, outreach, and student engagement opportunities to enhance research and better utilize DesignSafe's toolbox.

# NHERI COMMUNITY

https://www.araport.org

Search by names, identifiers

# ARAPORT
ARABIDOPSIS INFORMATION PORTAL

About | Data | Tools | Community | Developer | Help | Log in

The Araport API Manager infrastructure is undergoing an upgrade and maintenance on April 5th, 2016 from 09:00am to 6:00pm CDT. For more information, please see the **service disruption notice**. Please check back for a status update on the upgrade/maintenance outage. Thank you for your patience!

ALL | THALEMINE | JBROWSE

Search by names, identifiers, keywords ...

*Search ThaleMine, JBrowse and other site content*
*Search by identifiers or keywords e.g. AT4G19020 or epigenetic*

## New to Araport?

Araport is a one-stop-shop for *Arabidopsis thaliana* genomics. Araport offers gene and protein reports with orthology, expression, interactions and the latest annotation, plus analysis tools, community apps, and web services. Araport is 100% free and open-source. Registered members can save their analysis, publish science apps, and post announcements. READ MORE

REGISTER LOGIN

Araport11

## THALEMINE

## JBROWSE

## BLAST

## APP STORE

- https://bitbucket.org/agaveapi/cli
- Requires bash and python's json.tool
- Uses caching for authentication
- Parses JSON responses to condense output

- As a Linux user, this is home-sweet-home

# DEVELOPER COMMAND-LINE TOOLS

- Bleeding edge research will never be on a webpage
- Data exploration "outside the app" also needs to be captured
- An infrastructure for responsible computing at scale inevitably must support responsible data exploration
- Jupyter has broad OS support, domain adoption, domain libraries, and a more interactive UI

# WHAT ABOUT JUPYTER?

# AGAVEPY

- github.com/TACC/agavepy
- Pythonic wrapper for all Agave endpoints
- pip install agavepy
- Developers actively "dogfooding" the module
- (Obviously) usable within Jupyter
- Has had greater uptake by users (not just developers)

- Going one step further – give users a notebook
- jupyter.public.tenants.prod.agaveapi.co/
- (Free) account creation here: public.tenants.prod.agaveapi.co/create_account
- Beta implementation at the moment
  - data purges during updates
  - Limited capacity on the current VM
  - All notebooks run inside Docker containers

# AGAVE-AWARE JUPYTERHUB

- Full-featured developer portal
- Open-source reference implementation of an Angular Javascript portal built on Agave
- Additional Jupyter notebook examples
- Production-grade support for a hosted JupyterHub

# WHAT'S NEXT?


# THANKS!
# QUESTIONS?

Slides: tinyurl.com/FonnerSEA2016

Email: jfonner@tacc.utexas.edu

Twitter: @johnfonner

TACC: www.tacc.utexas.edu

Agave: www.agaveapi.co

AgavePy: github.com/TACC/agavepy