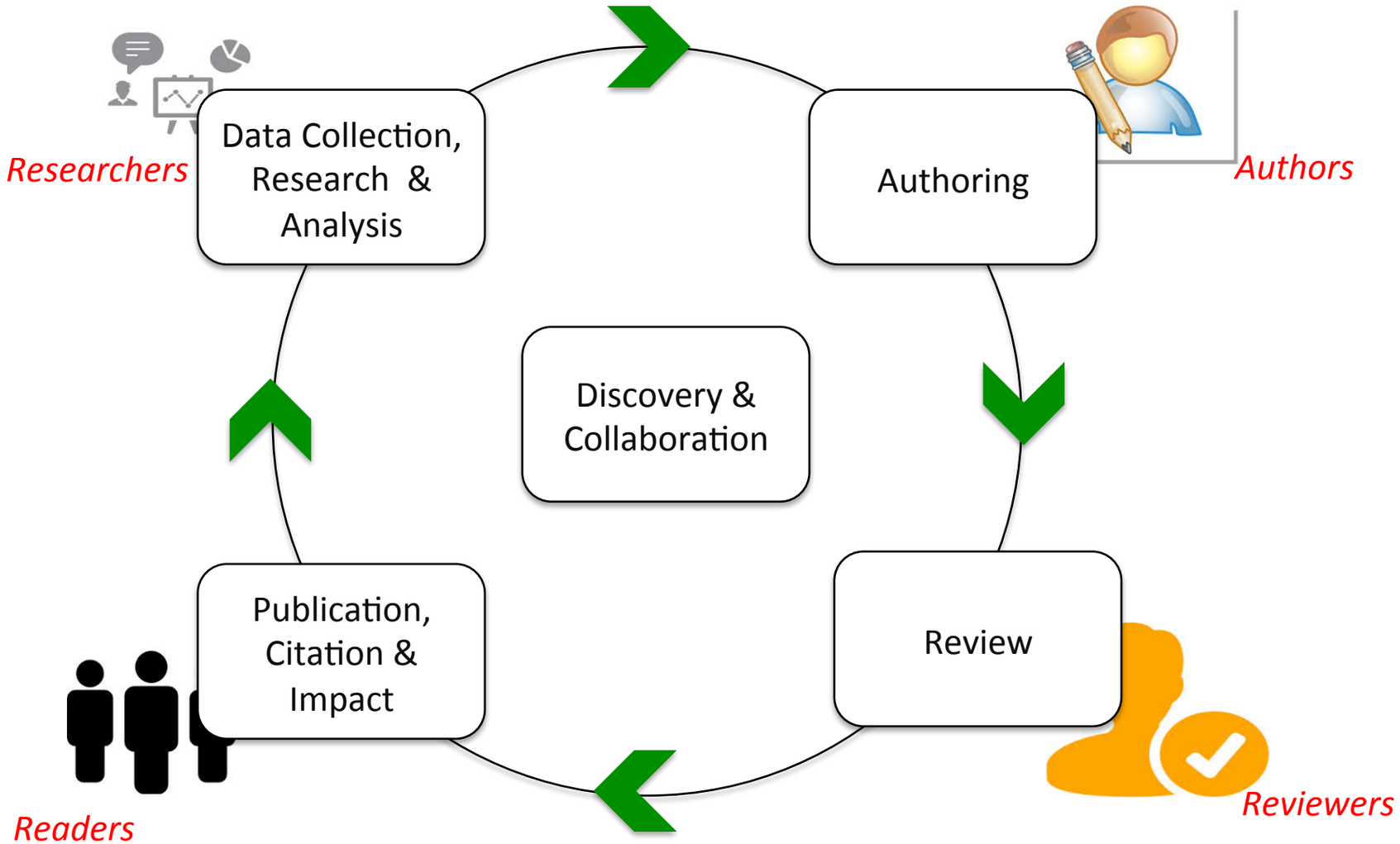
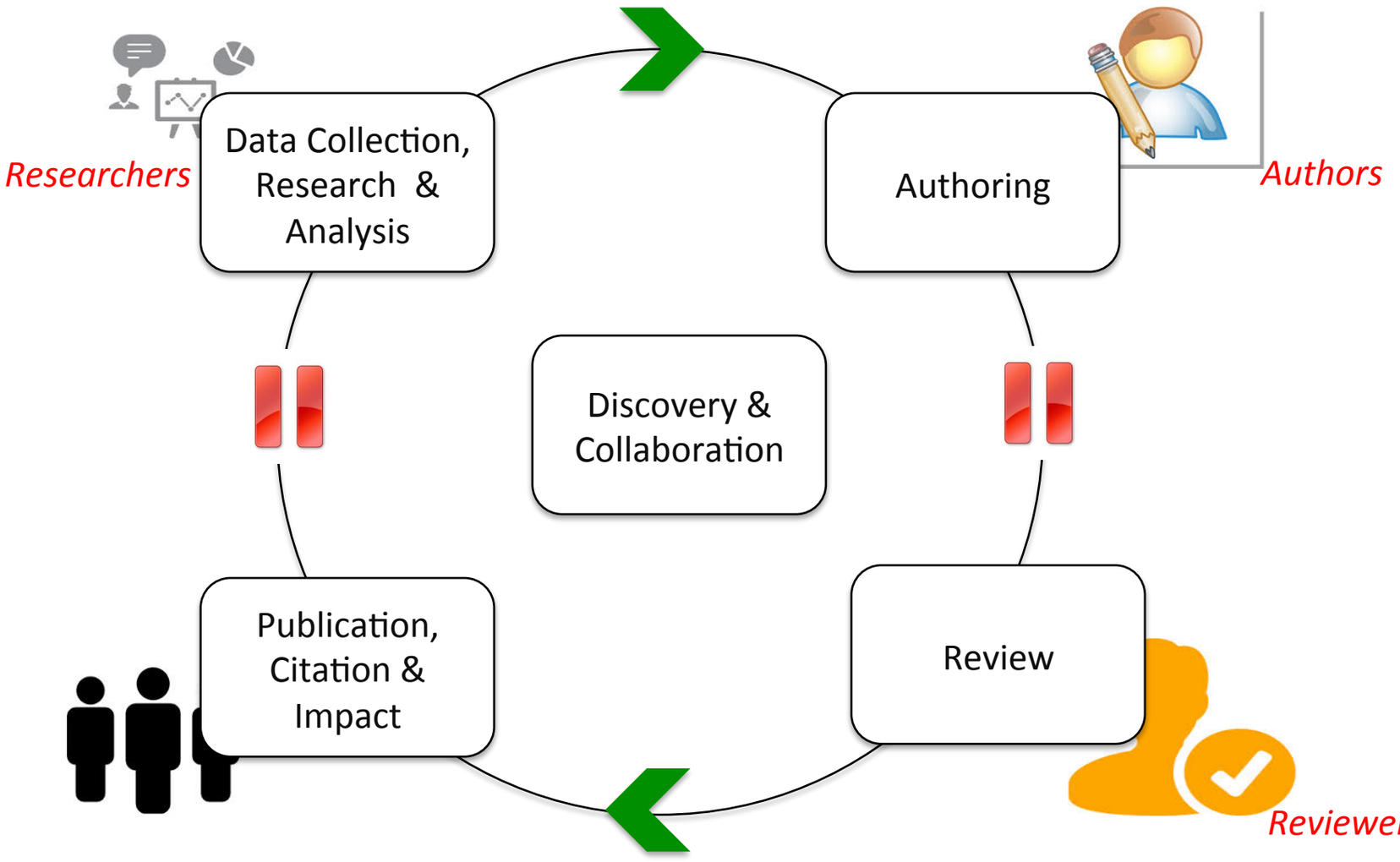
Science Dataspaces for Data Management and Reproducibility

Tanu Malik, Ian Foster, Kyle Chard
Jonathan Goodall, Scott Peckham, Joseph Baker, Mike Gurnis

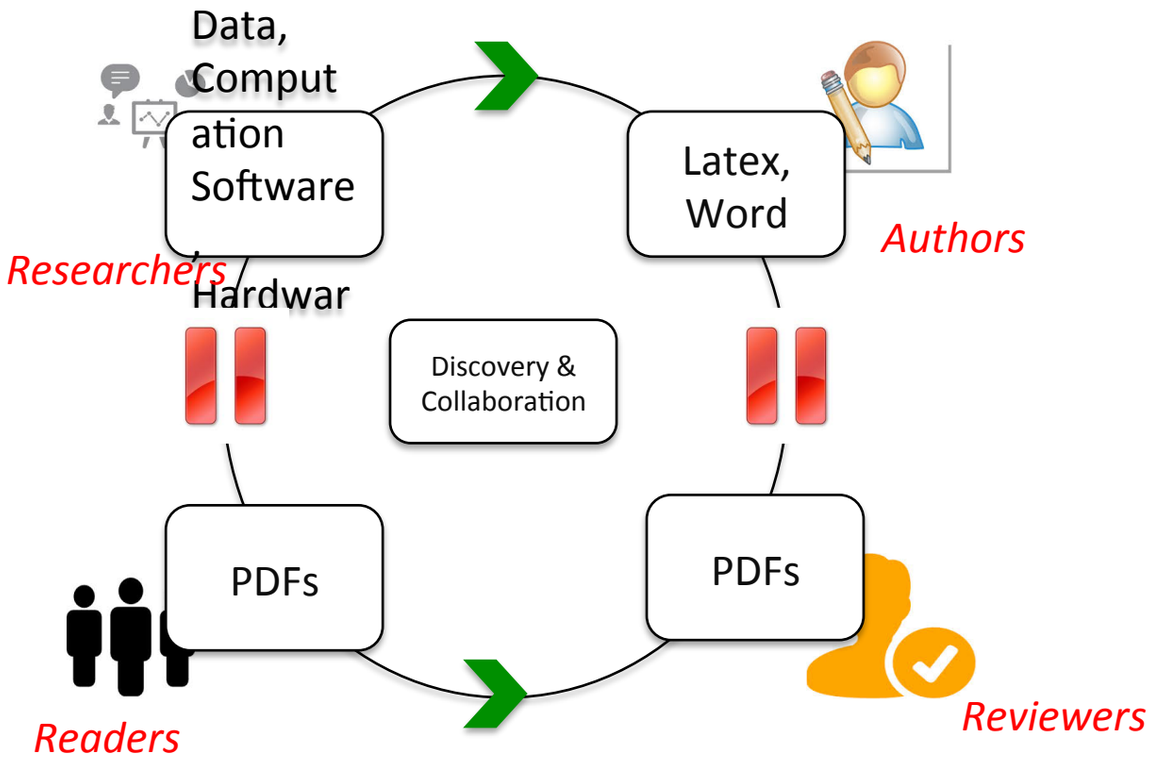
The scientific method is self-correcting



But...Scientific Publishing is Not



Computational Scientific Publishing is Broken



Computational science inputs are not linked with outputs.

- *Inputs:* Large quantities of data, complex data manipulation and/or numerical simulation use of large and often distributed software stacks, etc. (software, data, execution, environment)
- *Outputs:* Research papers (text-based, non-interactive)

Encourage Open-X

- X = access/code/data/design/standard
- Use the Internet and be more social

Dropbox

GitHub

figshare
credit for all your research

ckan The open source data portal software

DRYAD

F1000Research
OPEN SCIENCE • OPEN DATA • OPEN PEER REVIEW

(GIGA)n SCIENCE

altmetrics

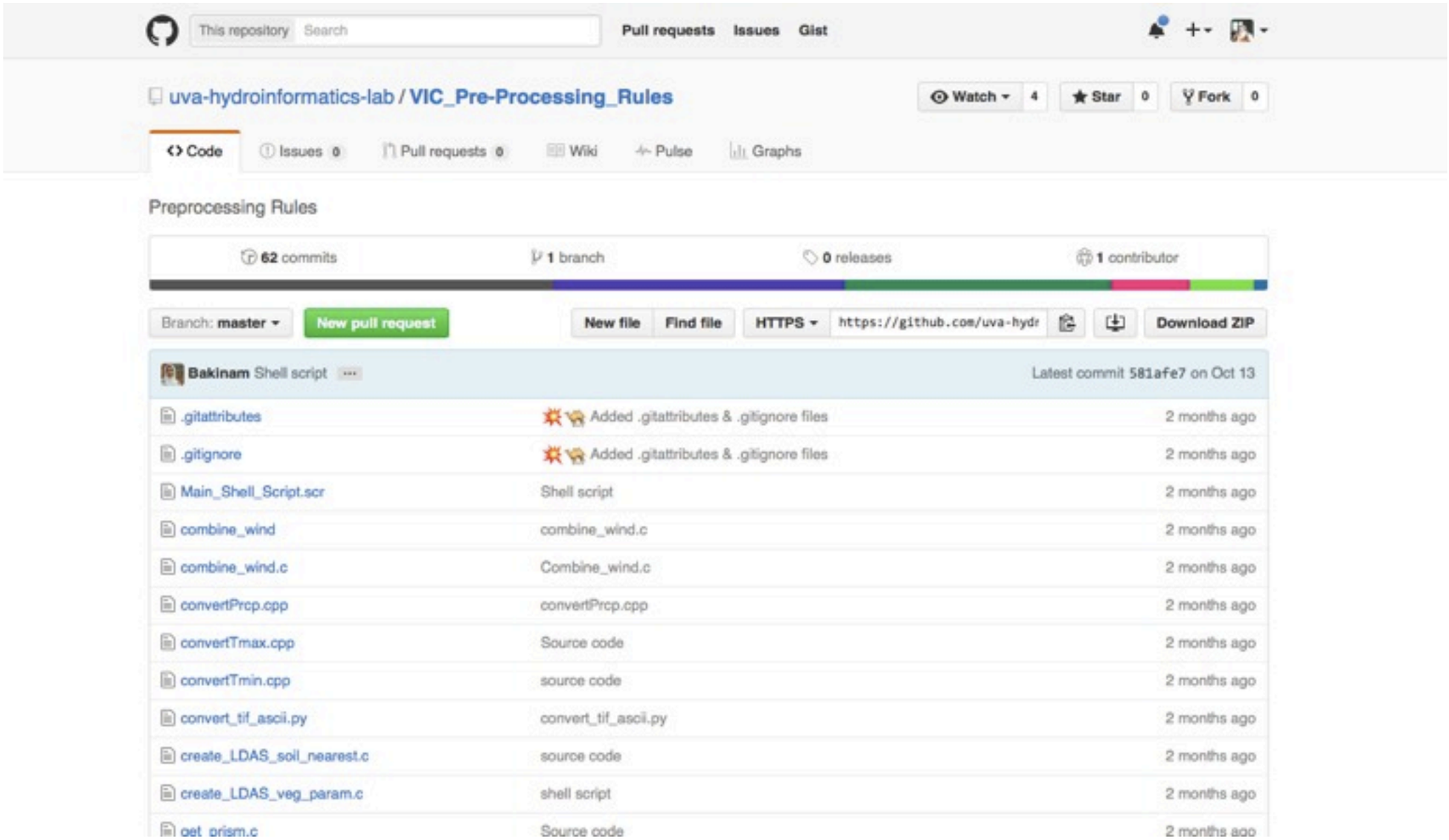
Two examples of Open-X

Share
with
collaborators

Share publicly
and
get social credit

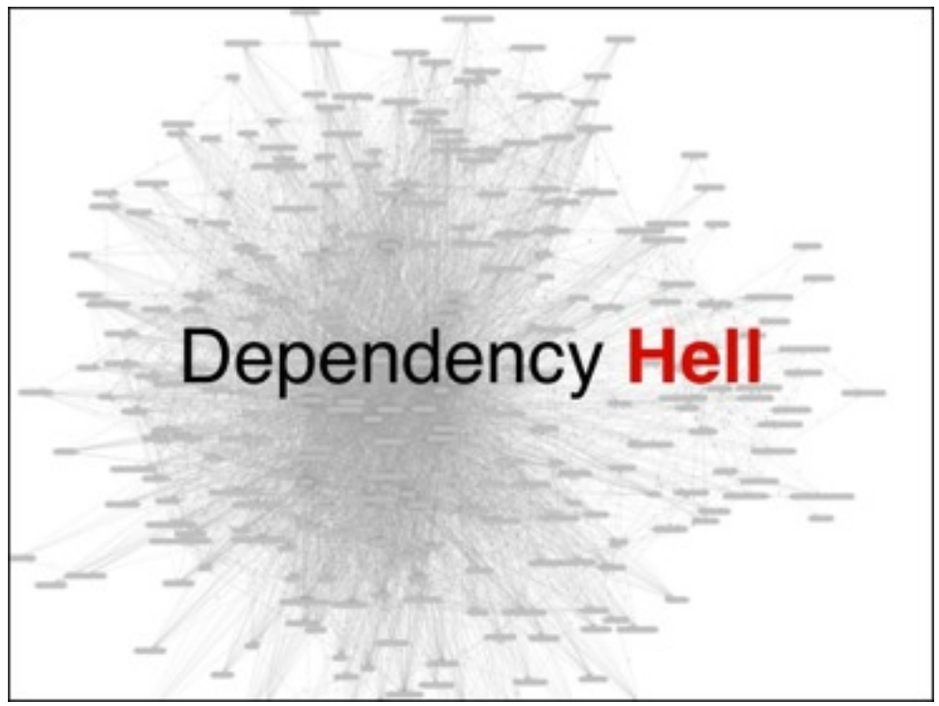
- Why aren't Github + DropBox sufficient for data management and reproducibility?

Shared Github Repo



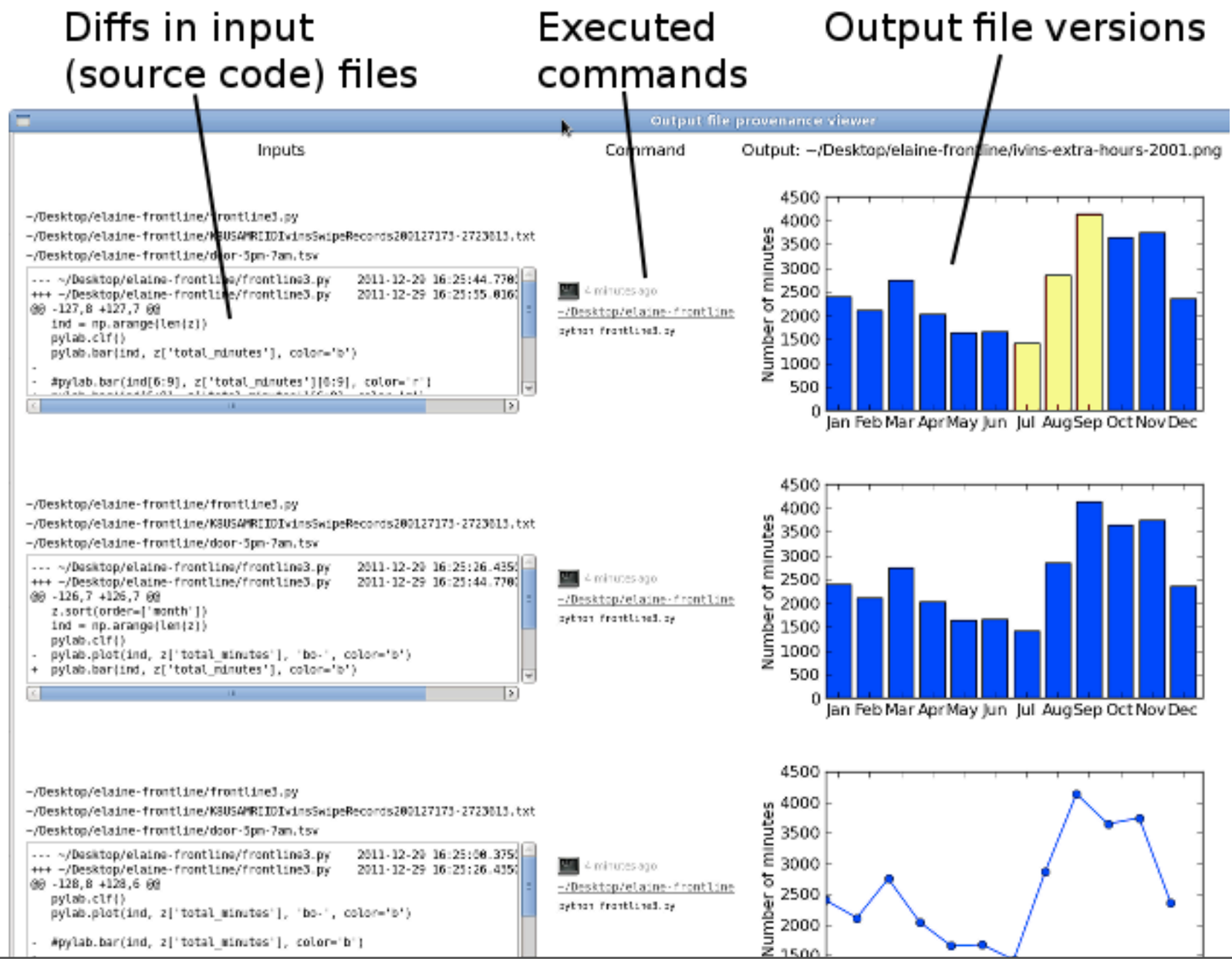
Missing Dependencies

- Has the author shared everything, including code, data, and environment?
  - Will this code, if downloaded, run in my computational environment?



Missing Lineage

- Which version of the data produced this result?



Missing Context

- What was I thinking when I produced this result?
  - Looking some documentation
  - Web-pages visited
  - Bash commands run
  - Associated tags




SciDataspace

Personalized, Shareable Dataspace for Data Management and Reproducibility

GeoDataspace

Log In | Sign Up

Username :

Username

Password :

Password

Sign in

GeoDataspace

The GeoDataspace framework is assists scientists and communities to create and maintain collections of geounits that pertain to a specific research project. To create and use GeoDataspace, please sign up and download the client.

EarthCube

NSF

GeoDataspace client

- Git and DropBox-like Python client for Linux and Mac OS X
  - annotate
    - provide semantic annotations
  - package
    - code, data, environment into Docker containers
  - track
    - tracks provenance of the scientific program

GDClient

Start the client

Available Commands

Annotate

Track

Package

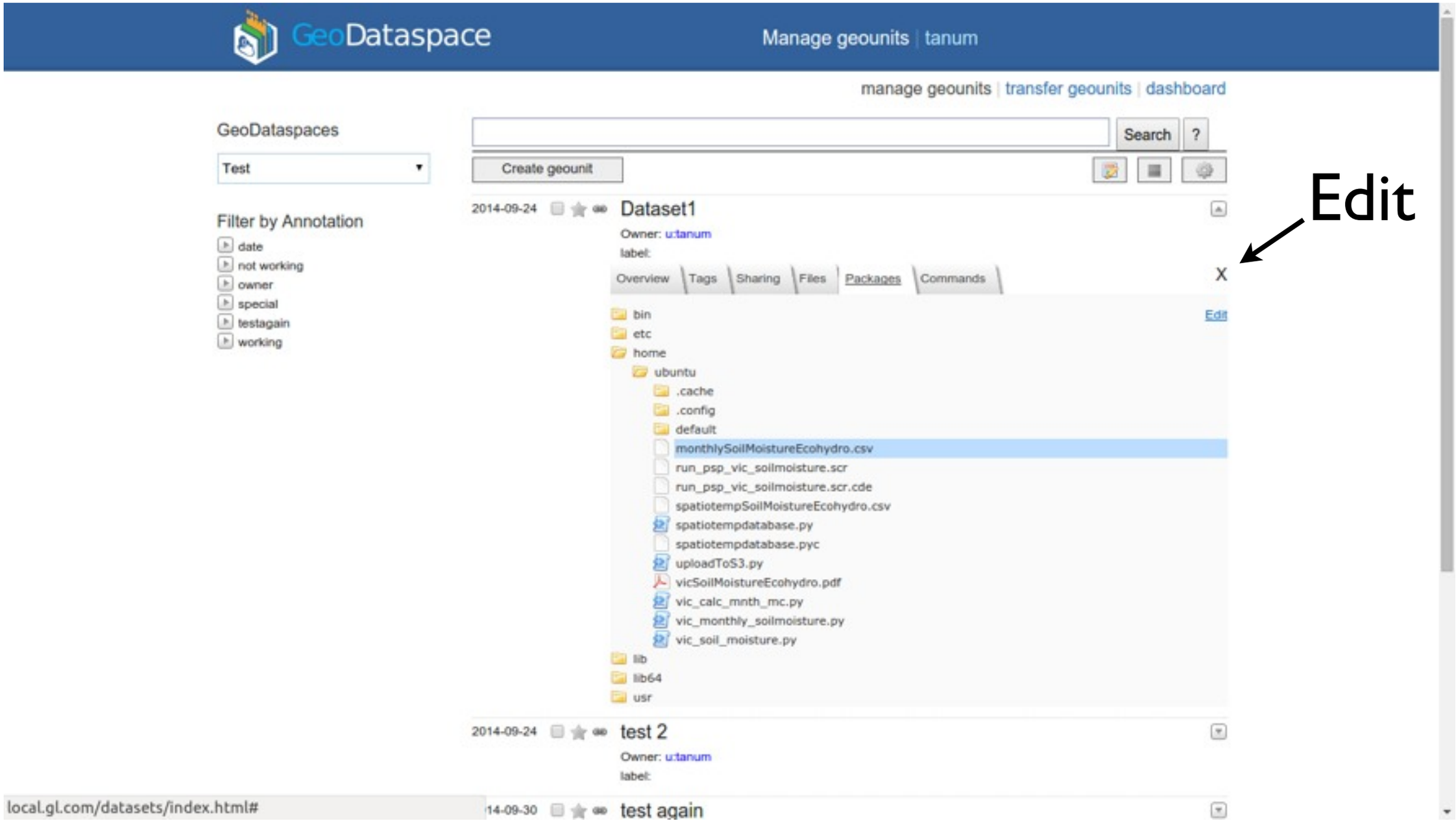
```
ubuntu@ip-172-30-1-16:~$ gdclient.py
enter "stop" to end session
gdclient commands start with  --
undefined > --
--add_member   --annotate     --geounit      --package      --stop         --track        --transfer
undefined > --geounit start AGU-Demo
AGU-Demo > --annotate geounit note:My_demo_geounit
annotation definition already exists
Added
AGU-Demo > --add_member /home/ubuntu/vicSoilMoistureEcohydro.pdf
80
AGU-Demo > --track ./run_psp_vic_soilmoisture.scr
   >  extracting data for soil moisture from VIC model simulation.....
   >  First python Script executed Succefully
   >  averaging monthly soil moisture from VIC model simulation .....
   >  Second python Script executed Succefully
   >  plotting  monthly soil moisture from VIC model simulation .....
   >  Third python Script executed Succefully
   >  Uploading Workflow output to S3 Bucket .....
   >  Finished!
0xf, 0x7ffccc0426a0, 0x2, 0x4000, 0x4, 0x51e7f, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x
0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x
0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0, 0x7f5d592853e0AGU-Demo >
AGU-Demo > --package
USAGE: package [provenance] list| delete| level [individual <program name>| collaboration <package id>| community]
AGU-Demo > --package
USAGE: package [provenance] list| delete| level [individual <program name>| collaboration <package id>| community]
AGU-Demo > --package
delete      level       list        provenance
AGU-Demo > --package list
9 packages available:
23e10c450fc08e1ba230ed0a2b66afbc250518b6    2015-10-16 16:27:07.992763     ./run_psp_vic_soilmoisture.scr
667c7d58b2fb1fae46bd92ab75333f34743718ec    2015-11-04 16:18:26.067837     ./run_psp_vic_soilmoisture.scr
3b3fe52344e918481fdb6d00483f54248e7f0810    2015-05-27 12:24:01.181754     ./run_psp_vic_soilmoisture.scr
8db88f041083a38d2f81c95a4516714e2473525f    2015-05-27 19:48:00.968017     ./run_psp_vic_soilmoisture.scr
4d8e2a80971daa4092c53203e70e732387c74d2a    2015-10-07 14:52:04.291792     ./run_psp_vic_soilmoisture.scr
1d19beeaf7ffe19d76410ecf7c0585fd44aa1c54    2015-05-27 20:10:12.716444     ./run_psp_vic_soilmoisture.scr
59cb817c9061e74814284c83c3895e9ac0e7dc5    2015-05-27 11:38:54.801533     ./run_psp_vic_soilmoisture.scr
7bc611bdcd93570f21a03d8a93581e6f163f9e62    2015-05-29 13:46:13.236476     /home/ubuntu/run_psp_vic_soilmoisture.scr
9978145896497ca3917be7d9fabb58ce734eb98a    2015-05-27 07:40:24.652386     ./run_psp_vic_soilmoisture.scr
AGU-Demo > --package level individual ./run_psp_vic_soilmoisture.scr
package_hash= 08a947f3fd4b57dfb7c44e21f05b2894af3d7615
AGU-Demo > --package level collaboration 08a947f3fd4b57dfb7c44e21f05b2894af3d7615
Successfully built image id  8834c910453d
Successfull
AGU-Demo > _
```

GeoDataspace Tools

Tool	Repeatability Feature	Users
SPADE[1]	Provenance (V)	*
CDE[2]	Packaging (P)	*
PTU[3]	P+V	
SciDataspace[4]	**P+V+Usability**	**Geoscience NSF EarthCube**
PTU-NFS[5]	P'+V' (in Network File System)	DOE High-Energy Physics
LDV: Light-weight Database Virtualization [6]	P'+V' (in databases)	Urban Science

1. https://github.com/ashish-gehani/SPADE
2. http://www.pgbovine.net/cde.html
3. https://gitlab.com/quanpt/provenance-to-use
4. https://bitbucket.org/tanum/scids-client

Preserved in GeoDataspace



Manipulate Geounits

Save

Graph

Packages

- bin
- etc
- home
  - ubuntu
    - .cache
    - .config
    - default
    - monthlySoilMoistureEcohydro.csv
    - run_psp_vic_soilmoisture.scr
    - run_psp_vic_soilmoisture.scr.cde
    - spatiotempSoilMoistureEcohydro.csv
    - spatiotempdatabase.py
    - spatiotempdatabase.pyc
    - uploadToS3.py
    - vicSoilMoistureEcohydro.pdf
    - vic_calc_mnth_mc.py
    - vic_monthly_soilmoisture.py
    - vic_soil_moisture.py
- lib
- lib64
- usr

/home/karthik/Desktop/Extra_Files/FLP/Tanu/Packages_tmp/667c7d58b2fb1fae46bd92ab75333f34743718ec/cde-package/cde-root/home/ubuntu/monthlySoilMoistureEcohydro.csv

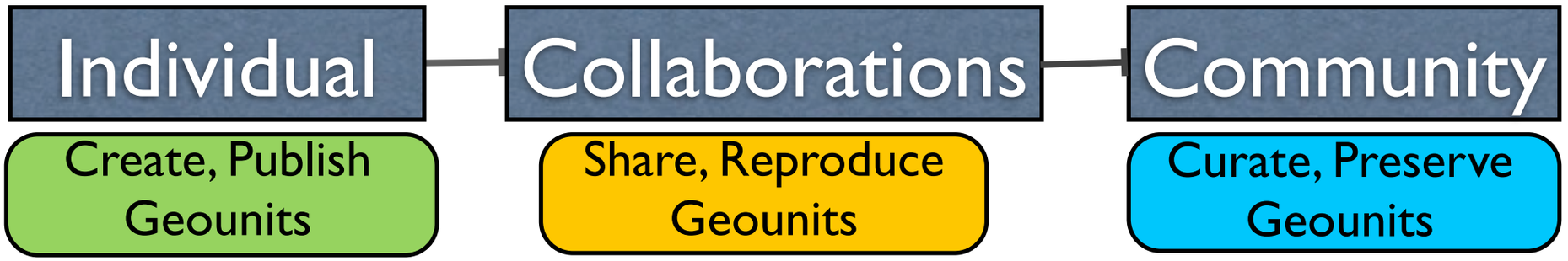
```
 98  mc,200601,0.23,0.24,0.28
 99  mc,200602,0.22,0.23,0.29
100  mc,200603,0.20,0.21,0.27
101  mc,200604,0.19,0.19,0.23
102  mc,200605,0.21,0.20,0.18
103  mc,200606,0.21,0.21,0.15
104  mc,200607,0.20,0.20,0.13
105  mc,200608,0.21,0.20,0.11
106  mc,200609,0.21,0.23,0.16
107  mc,200610,0.20,0.21,0.14
108  mc,200611,0.23,0.24,0.19
109  mc,200612,0.21,0.22,0.24
110  mc,200701,0.23,0.24,0.29
111  mc,200702,0.22,0.23,0.31
112  mc,200703,0.21,0.22,0.32
113  mc,200704,0.21,0.21,0.27
114  mc,200705,0.19,0.21,0.22
115  mc,200706,0.20,0.17,0.15
116  mc,200707,0.20,0.17,0.12
117  mc,200708,0.16,0.13,0.10
118  mc,200709,0.18,0.14,0.10
119  mc,200710,0.18,0.14,0.09
120  mc,200711,0.18,0.17,0.10
121  mc,200712,0.21,0.20,0.10
122
```

Shell Access:

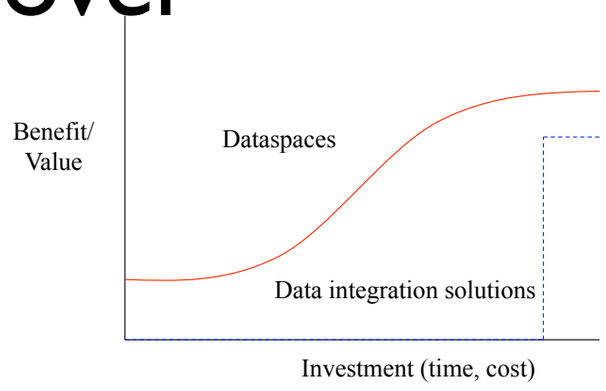
```
karthik@murray:~$
karthik@murray:~$
```

Track the workflow

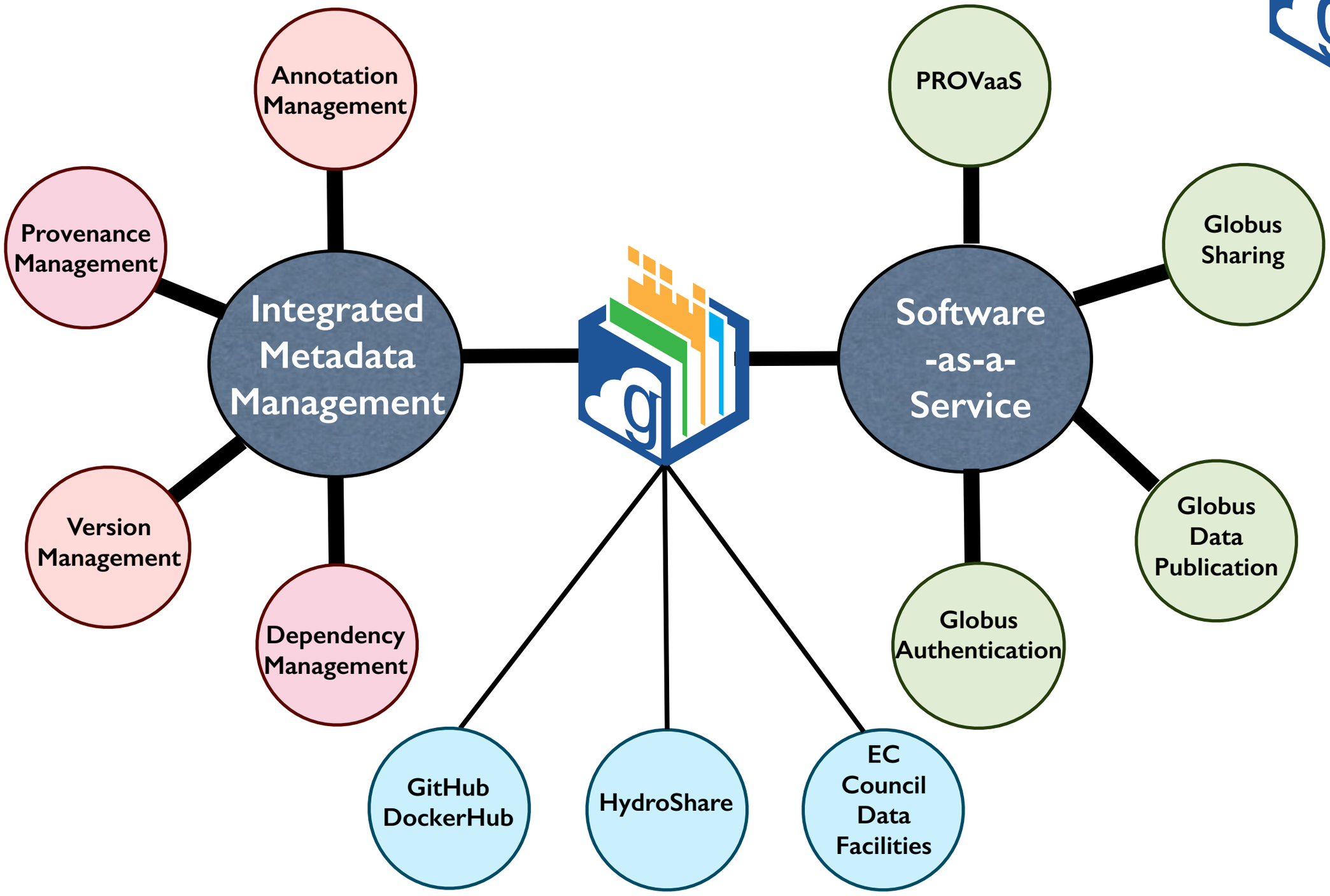
Preserved in a Dataspace



- Cloud-hosted Dataspace that can be shared with the collaboration with the help of services, and becomes standardized over time.

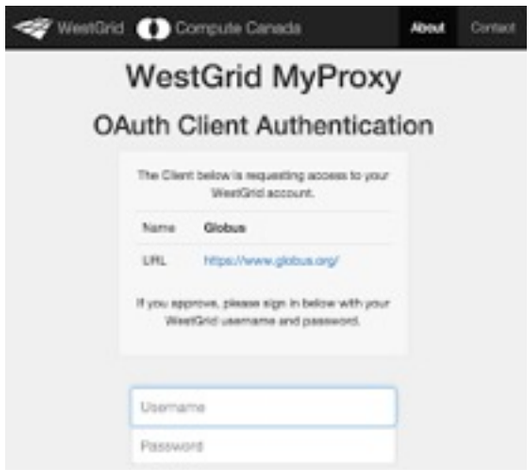


GeoDataspace Architecture

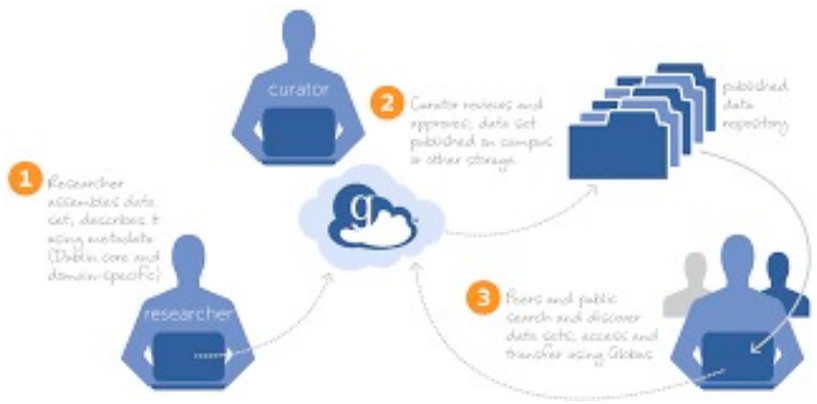


Globus Services

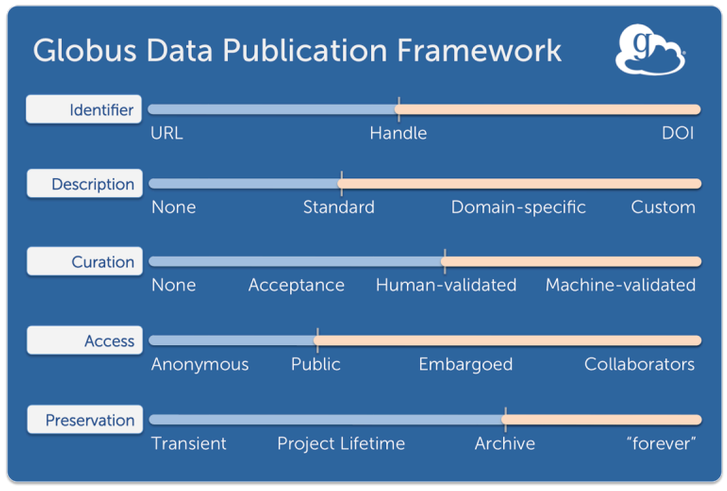
Authentication



Transfer

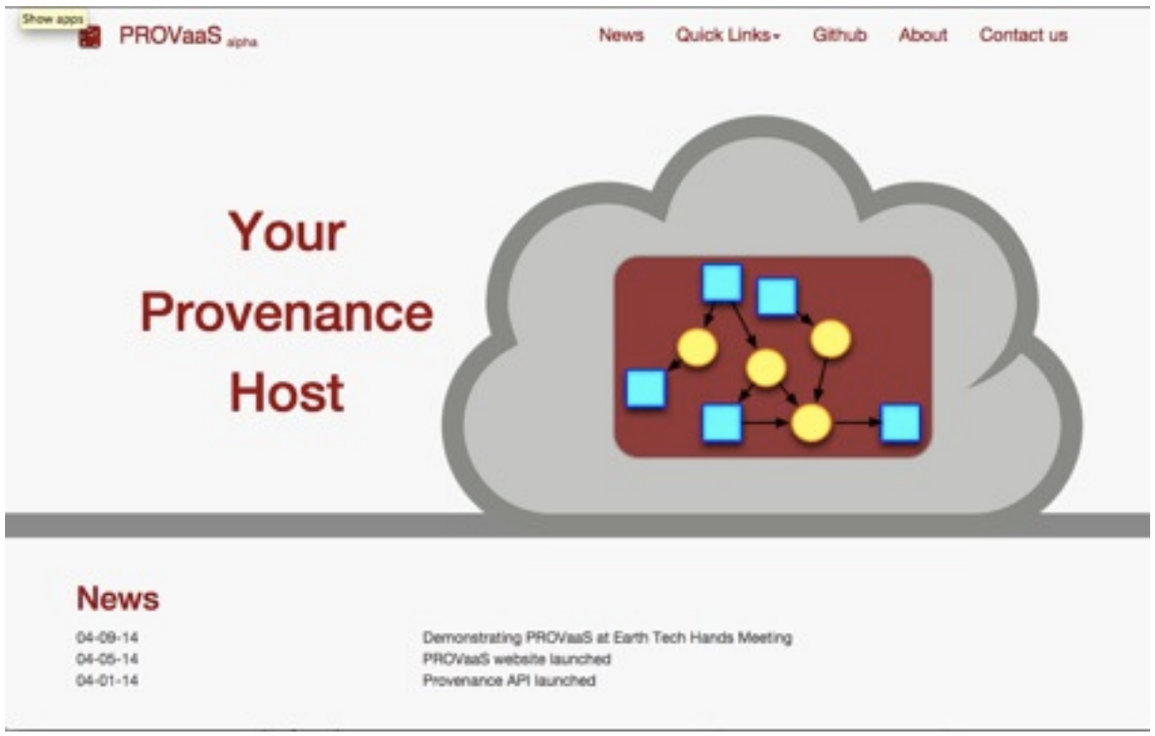


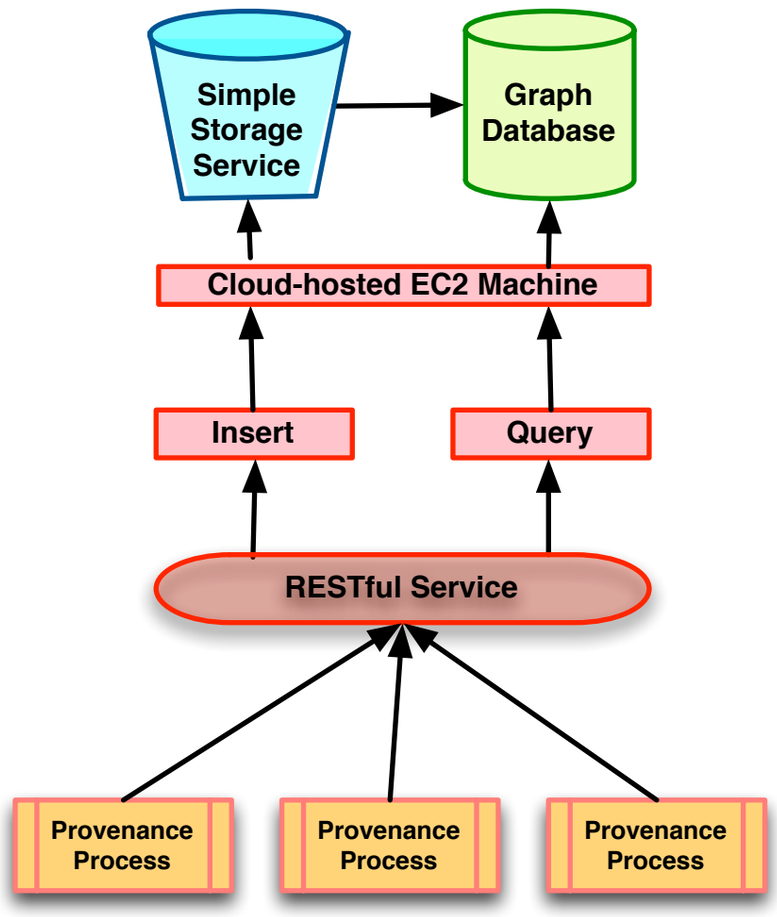
Sharing



Publication

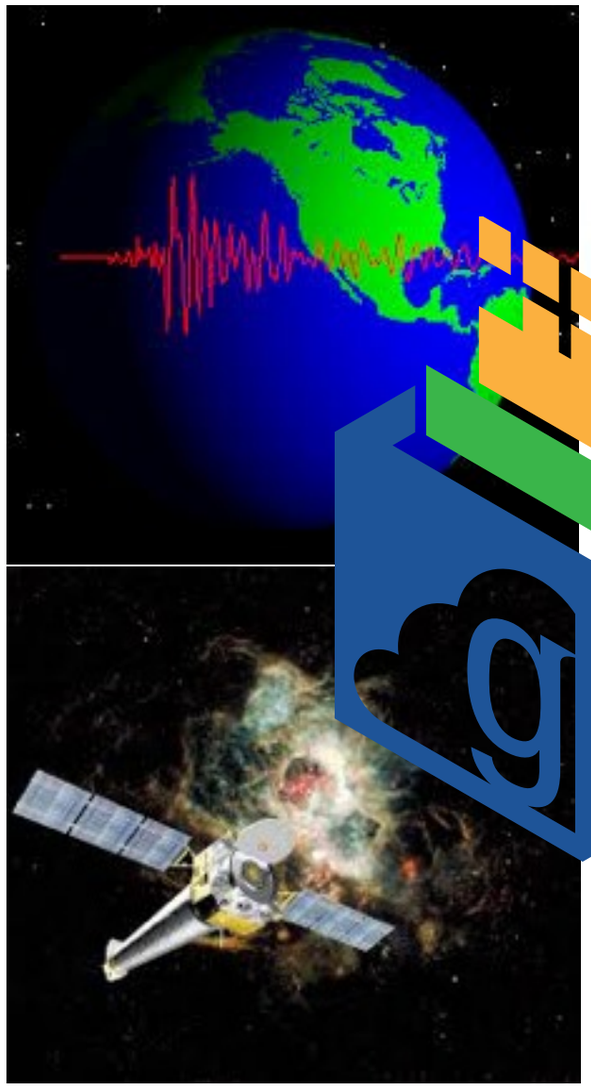
PROVaaS


Show apps
PROVaaS alpha
News
Quick Links
Github
About
Contact us
Your
Provenance
Host
News
04-09-14
04-05-14
04-01-14
Demonstrating PROVaaS at Earth Tech Hands Meeting
PROVaaS website launched
Provenance API launched


Simple Storage Service
Graph Database
Cloud-hosted EC2 Machine
Insert
Query
RESTful Service
Provenance Process
Provenance Process
Provenance Process

Bringing Order to Chaos in Science Use Cases

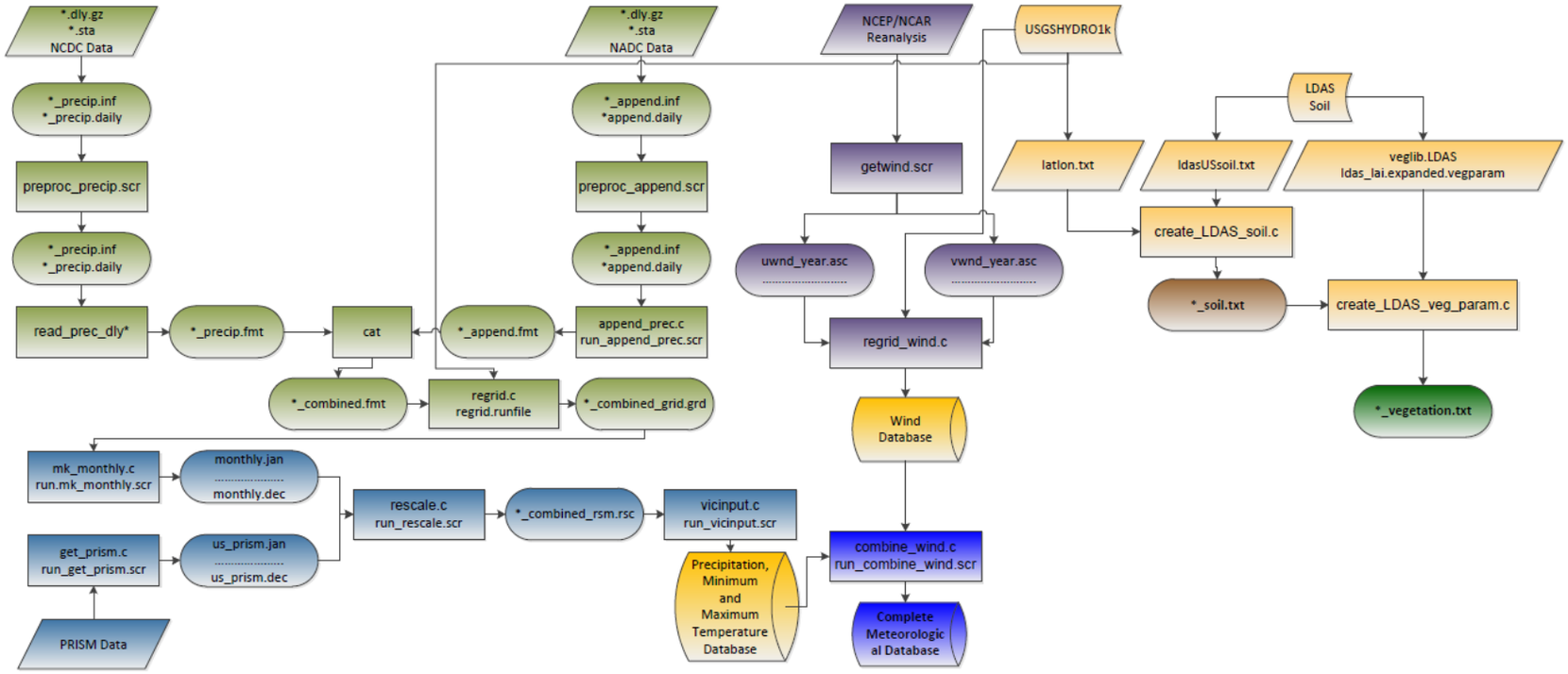
Solid Earth Hydrology



Space Science CSDMS

Hydrology

(Curation, Encapsulation, Reuse)



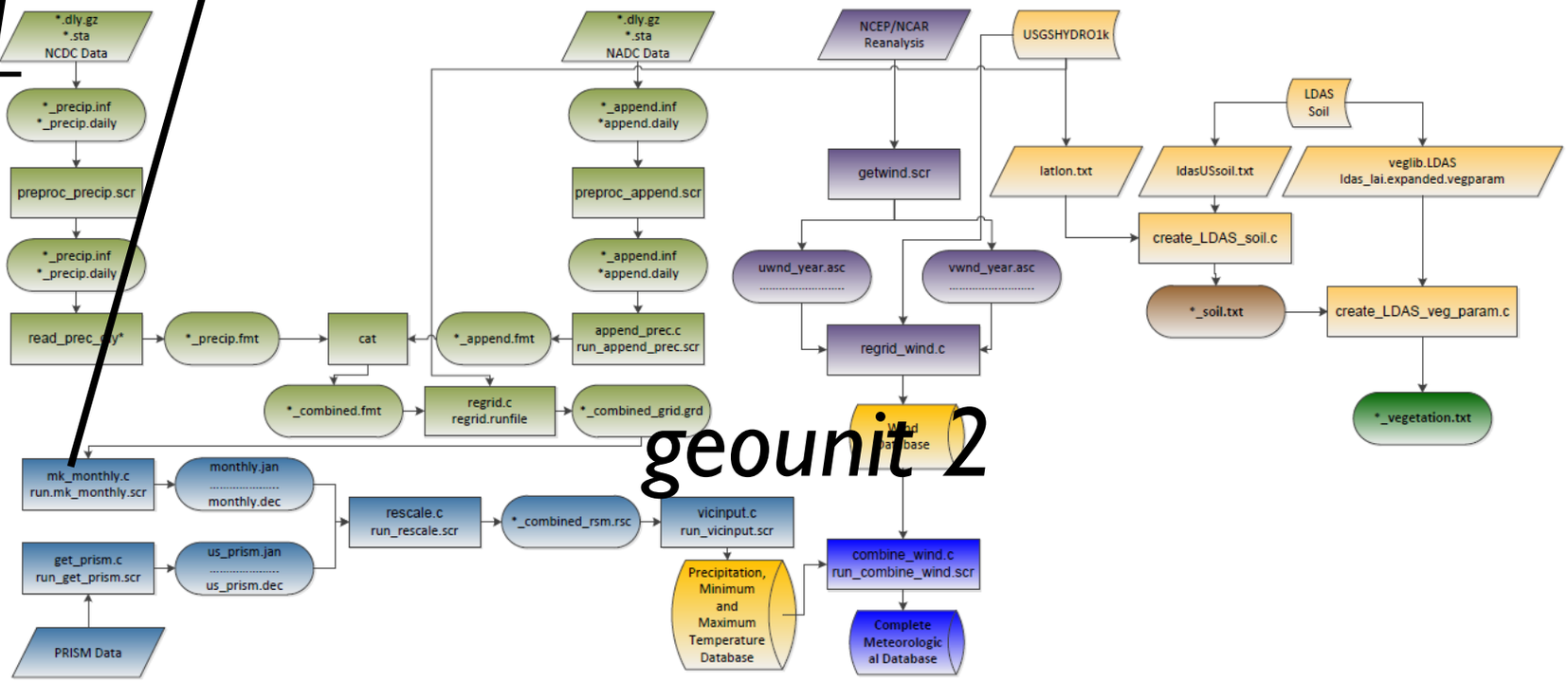
- Data processing steps for the VIC model

The reality

```
irods@ec2-54-86-215-185:~/iRODS/server/bin$ cd cmd/
irods@ec2-54-86-215-185:~/iRODS/server/bin/cmd$ ls
adjprcp_rsm.rsc                 forcingData                     old                             rescale_ir.scr                  stationinfo.txt
adjtmin_grid.grd                gauss_t62_lat.list              osgeo.bak                       run_combine_wind_ir.scr         temp_wind
amqprecv.py                     gauss_t62_lon.list              output_1.txt                    run_combine_wind.scr            TerraPopCountieswzVicOutput.py
amqpsend.py                     gdal.py                         PamAuthCheck                    run_convert_prcp.scr            test
Basin                           gdal.pyc                        PopulationVsSoilMoisture.scr    run_convert_tif_ascii.scr       test3.py
basin_prcp_adj.fmt              get_prism                       popvssm.run                     run_convert_tmax.scr            test_execstream.py
basin_prcp.fmt                  get_prism.c                     Prcp                            run_convert_tmin.scr            test.py
boundaries_US_SLAD_2010.dbf     get_prism_ir.scr                prcp.daily                      run_get_prism.scr               test.scr
boundaries_US_SLAD_2010.prj     getwind.scr                     prcp.inf                        run.mk_monthly.scr              tiff2ascii.py
boundaries_US_SLAD_2010.sbn     hello                           prcp_tobAdj.scr                 run_psp_vic_evapotranspiration.scr Tmax
boundaries_US_SLAD_2010.sbx     hello.scr                       prec_tob_adj                    run_psp_vic_sm.scr              tmax_tob_adj
boundaries_US_SLAD_2010.shp     inputPrcp.scr                   prec_tob_adj.f                  run_psp_vic_soilmoisture1.scr   tmax_tob_adj.f
boundaries_US_SLAD_2010.shp.xml inputTmax.scr                   prec_tob_adj.input              run_psp_vic_soilmoisture2.scr   tmax_tobAdj.scr
boundaries_US_SLAD_2010.shx     inputTmin.scr                   preproc_precip.scr              run_psp_vic_soilmoisture3.scr   Tmin
build                           irodsAgent                      prism                           run_psp_vicSoilMoistureComparison.scr tmin_tob_adj
catchment.dbf                   irodsReServer                   prism-rawdata                   run_psp_vic_soilmoisture.scr    tmin_tob_adj.f
catchment.prj                   irodsServer                     python_script1.py               run_regrid_wind_ir.scr          tmin_tobAdj.scr
catchment.shx                   irodsServerMonPerf              raw_wind                        run_regrid_wind.scr             tob
checkData                       irodsXmsgServer                 read_prec_dly                   run_rescale.scr                 univMSSInterface.sh
climate                         latlong99.txt                   read_prec_dly.f                 run_vicinput_ir.scr             uploadToS3.py
cmd                             latlon.txt                      readRodsLog.py                  run_vicinput.scr                vacuumdb.pl
combine_wind                    LDAS                            read_temp_dly.f                 script_1.scr                    vegetation
combine_wind.c                  ldas_lai.expanded.vegparams     read_tempn_dly                  script1.scr                     vic_calc_mnth_mc1.py
convertPrcp                     ldas_latlon2.scr                read_tempx_dly                  script_2.scr                    vic_calc_mnth_mc2.py
convertPrcp.cpp                 ldas_latlon3.scr                regrdPrcp                       script_3.scr                    vic_calc_mnth_mc.py
convert_tif_ascii.py            ldas_latlon4.scr                regrd_prcp.scr                  script_4.scr                    vic_calc_mnth_peb.py
convertTmax                     ldas_latlon.scr                 regrdTmax                       script_5.scr                    vic_evapotranspiration.py
convertTmax.cpp                 ldas_soil.scr                   regrd_tmax.scr                  script_6.scr                    vicinput
convertTmin                     ldas_veg.scr                    regrdTmin                       script_7.scr                    vicinput.c
convertTmin.cpp                 LeastSoilMoistureduration.py    regrd_tmin.scr                  script.save                     vic_monthlyPEBestimates.py
coop_tob.his                    list.pl                         regridPrcp                      script.scr                      vic_monthly_soilmoisture2.py
create_LDAS_soil_nearest        mask2latlon                     regridPrcp.runfile              script_test.scr                 vic_monthly_soilmoisture.py
create_LDAS_soil_nearest.c      mask2latlon.c                   regridTmax                      script_vic_pre-processing.scr   VIC_Pre_processing_script.scr
create_LDAS_veg_param           metadata.txt                    regridTmin                      sm_comparison.py                vic_soil_moisture3.py
create_LDAS_veg_param.c         meteoCombined                   regridTmin.runfile              sm_seasonal.py                  vic_soil_moisture.py
data_US_SLAD_2010.csv           mk_mnth                         regrid_wind                     smseasonal.scr                  vic_spatiotempdatabase.py
default                         mk_monthly                      regrid_wind.c                   soil                            wind_latlong.txt
DEM.tif                         mk_monthly.c                    rescale                         spatiotempdatabase.py           write.py
DSMwithpopulation.py            mk_monthly_ir.scr               rescale.c                       spatiotempdatabase.pyc
```

Organized in space

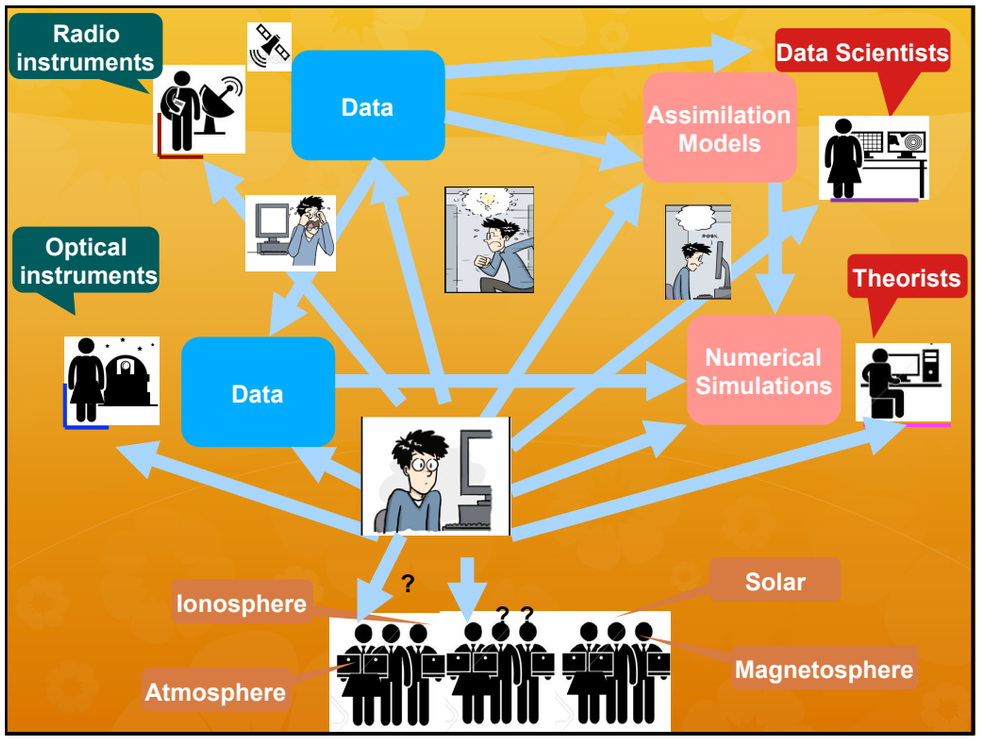
geounit 1

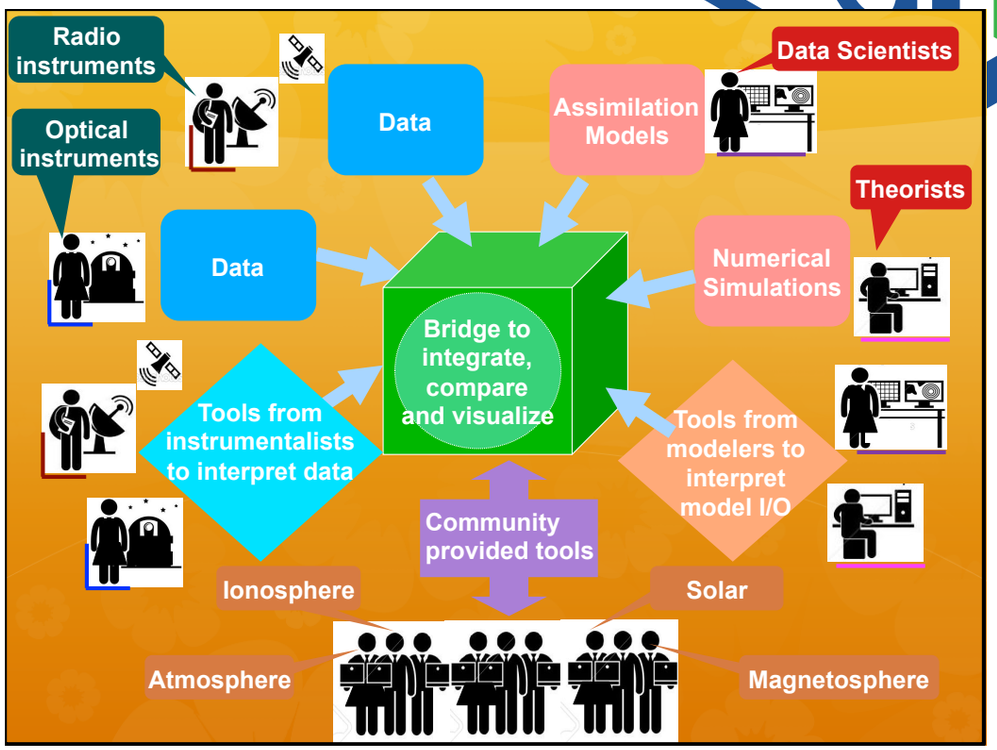


geounit 2

Space Science

(Model and Data Integration)





InGeo: Integrative GeoScience Observatory

EarthCube Integrative Activity: Asti Bhatt, Russell Cosgrove (SRI International)

Summary

- Capability-rich tools
  - Self-curation
  - Tracking cause and effect between data and models
  - Ensuring reproducibility
  - Support services: discovery, sharing, publication
- **Simplify** model and data management for computational/data geoscientists
- **Faster and reproducible hand-shakes** among faculty and students
- **Cuts down IT support** for modeling centers and real science

Track it!

- Science Usecases, Reports, Presentations, News
  - http://workspace.earthcube.org/geodataspace
- Source Code (Public Release Pending)
  - http://github.com/TanuMalik/SciDataspace/Geo
- The GeoDataspace (Internal URL)
  - https://scidataspace.org/geo

Acknowledgements

- National Science Foundation
- EarthCube Community
- Globus team
- CI team