An Extensible Geospatial Data Framework (GeoEDF) for FAIR Science

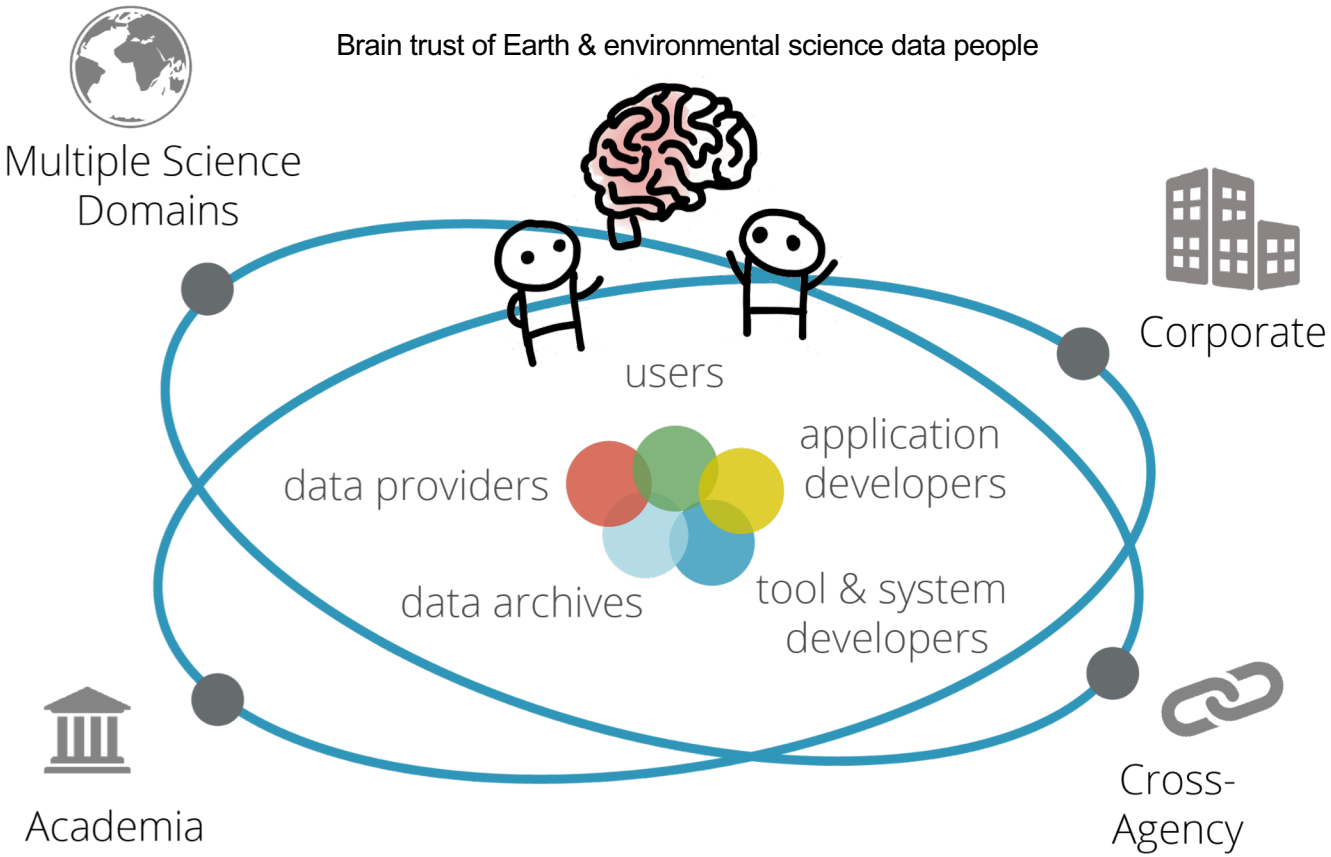
Dr. Carol Song
Sr. Research Scientist, Purdue Univ.

Assisted by: Rajesh Kalyanam
Research Scientist, Purdue Univ.

Data to Action Webinar: Increasing the Use and Value of Earth Science Data and Information

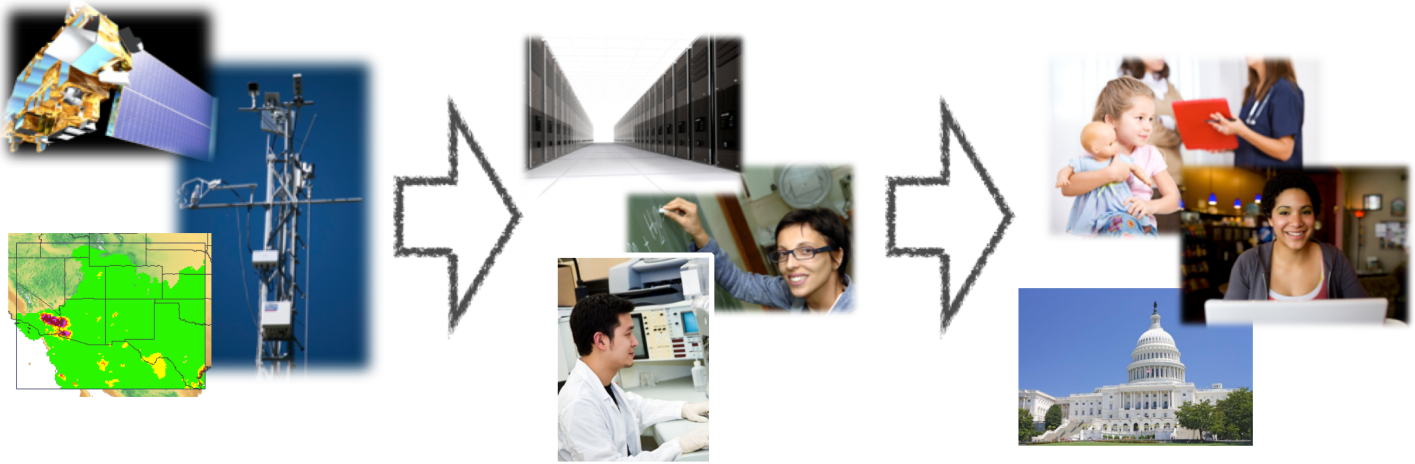
October 25th, 2019 | 1:00 pm ET

ESIP COMMUNITY



ESIP Vision

To be a leader in promoting
*the **collection, stewardship and (re)use***
Of Earth science data, information and knowledge
that is responsive to societal needs.



An Extensible Geospatial Data Framework (GeoEDF) for FAIR Science

Dr. Carol Song
Sr. Research Scientist, Purdue Univ.

Assisted by: Rajesh Kalyanam
Research Scientist, Purdue Univ.

Data to Action Webinar: Increasing the Use and Value of Earth Science Data and Information

October 25th, 2019 | 1:00 pm ET

NSF AWARD #1835822

An Extensible Geospatial Data Framework (GeoEDF) for FAIR Science

Carol Song, Rajesh Kalyanam, Purdue University

ESIP “Data to Action” Webinar
OCTOBER 25, 2019

The GeoEDF Project

An Extensible Geospatial Data Framework Towards FAIR Science

To help data-driven sciences to be more
Findable, Accessible, Interoperable, Reusable

funded by NSF CSSI program (Cyberinfrastructure for Sustained Scientific Innovation), Data Framework track, $4.5M

October 2018 - September 2023

Project Leadership

Jian Jin

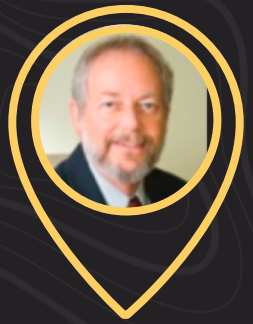
Plant phenotyping & sensors

Venkatesh Merwade

Flood modeling & visualization

Carol Song

PI
Cyberinfrastructure



Jack Smith

Water Quality & resource management

Uris Baldos

Sustainable development

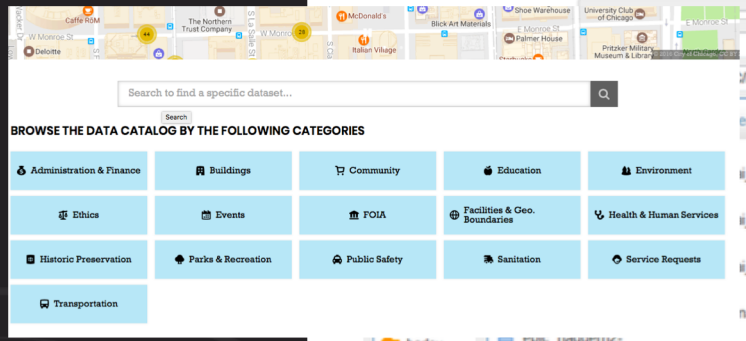
OVERVIEW

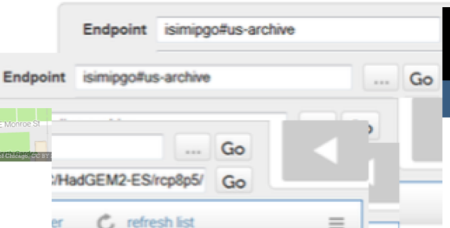
Extensible, Geospatial Data Framework Towards FAIR Science (GeoEDF)

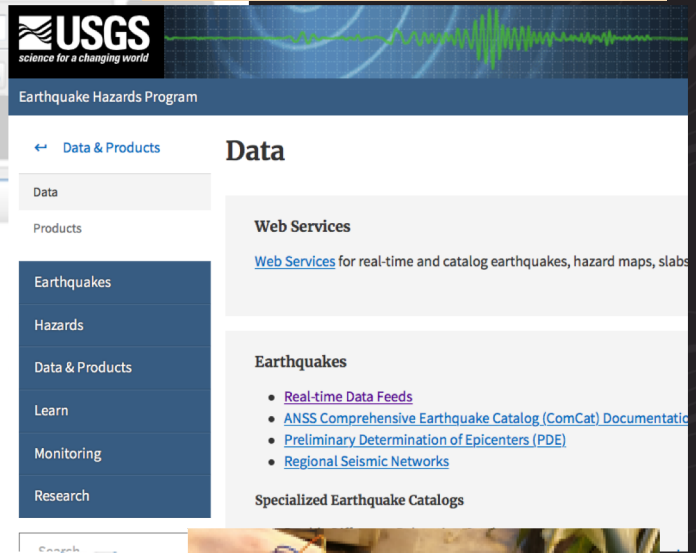
Rapid growth of geospatial, geo-referenced data

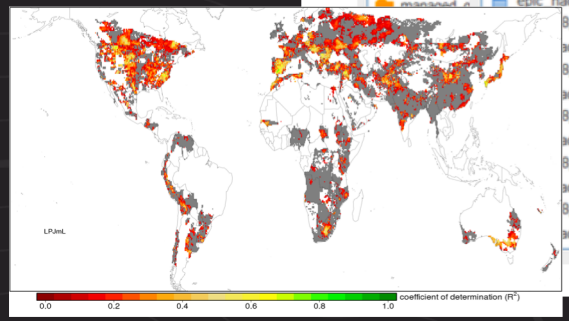
Data made public by cities

Hazard data provided by government





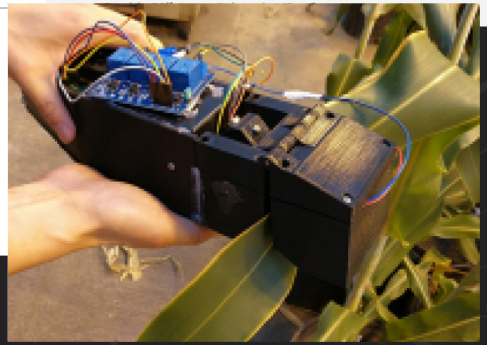




Modeling output data

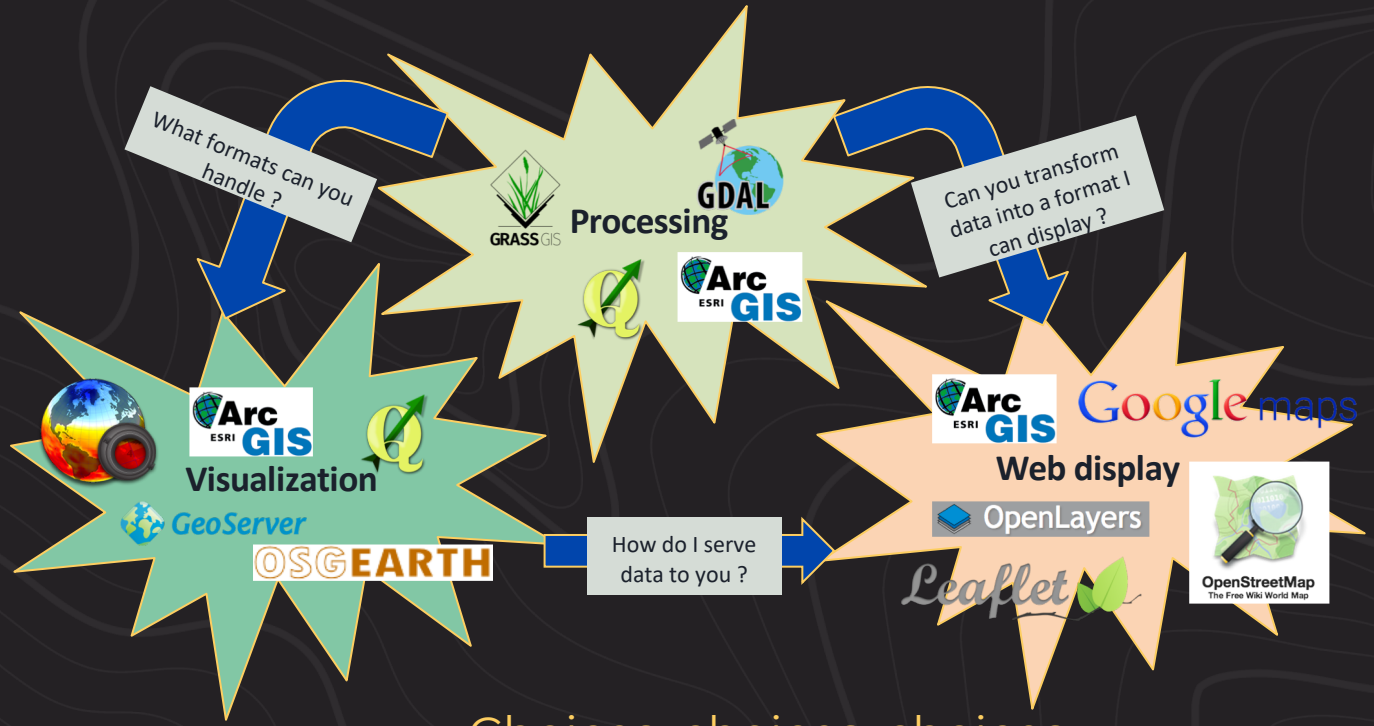
Remote sensing data

Data collected in the field



Software stack for spatial data

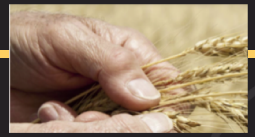
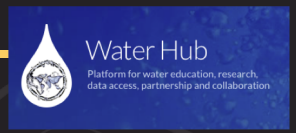
It is definitely not trivial to deal with geospatial data (processing, displaying, exchange/sharing, etc)

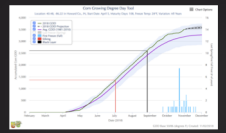


Choices, choices, choices

Take 1: GABBs -- Geospatial Data Building Blocks

Geospatial Data Building Blocks (GABBs)


. . .

Integrated data management environment with **built-in** geospatial data support

Data visualization builders and tools that require **no programming**

Publication of data and tools (DOI)

Toolkits for rapid application development, **no GIS programming expertise** required

Data service **API**, interoperability

Easy to use and replicate

hubzero+

G·A·B·B·S
geospatial data analysis building blocks

iRODS

Geoserver

PostGIS

Solr

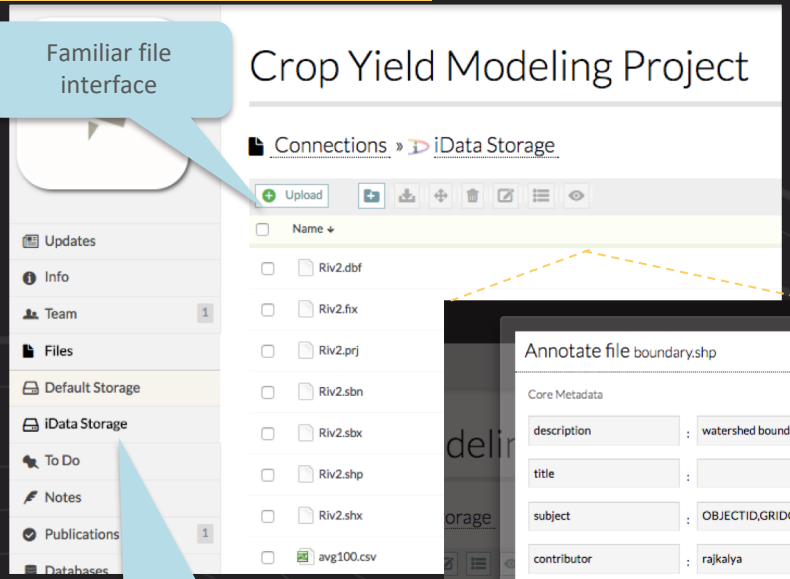
globus online

Community Clusters

XSEDE
Extreme Science and Engineering Discovery Environment

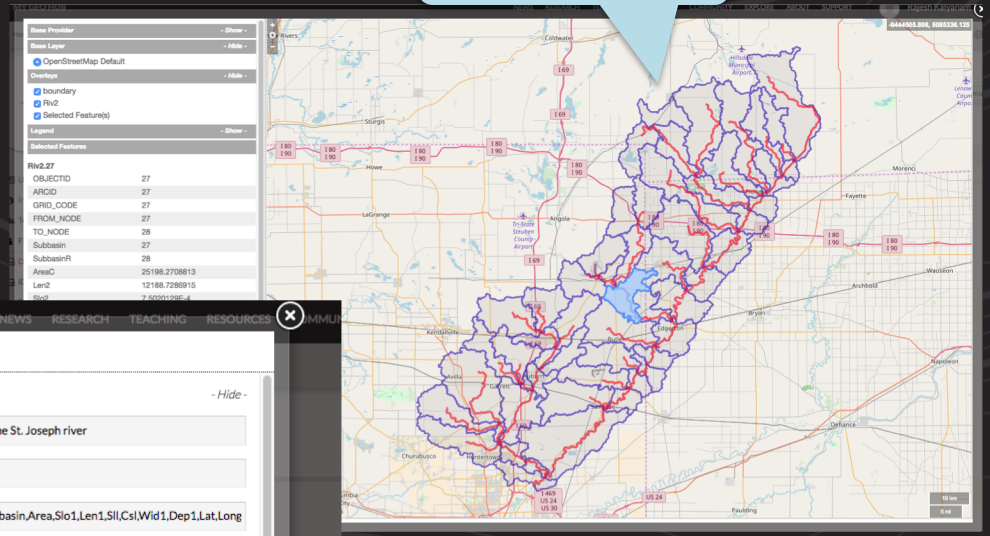
Integrated Geospatial Data Platform

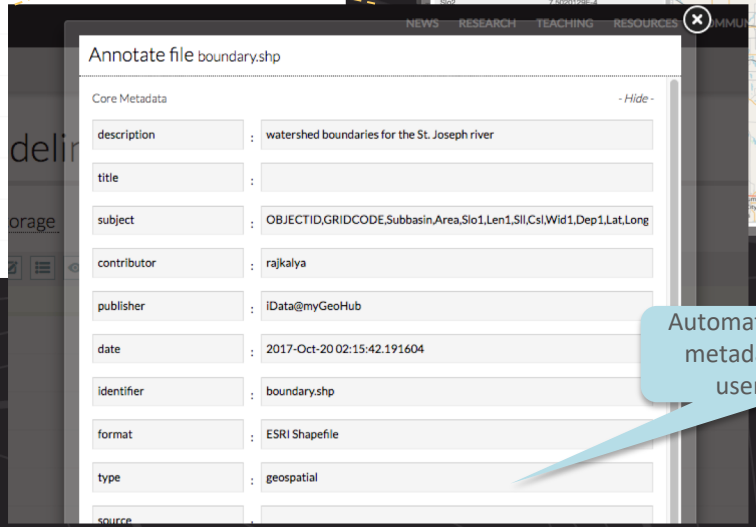
Familiar file interface



Storage backed by iRODS data management

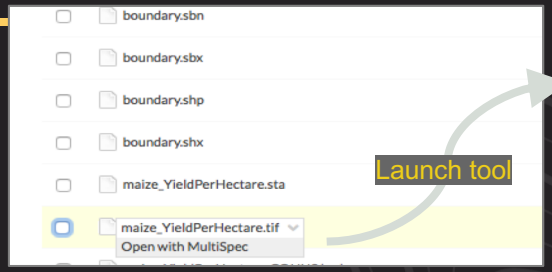
Preview and explore geospatial files right in the browser





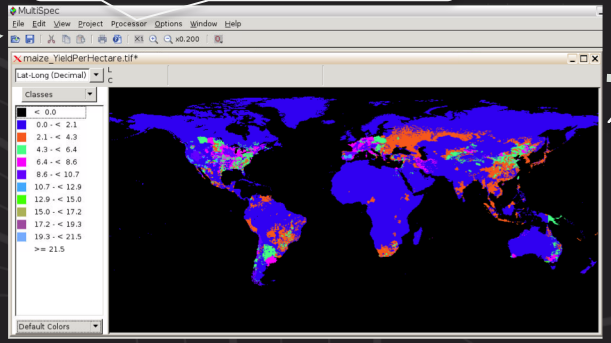
Automatically extracted metadata on upload, user extensible

Example workflow in one place



Launch tool

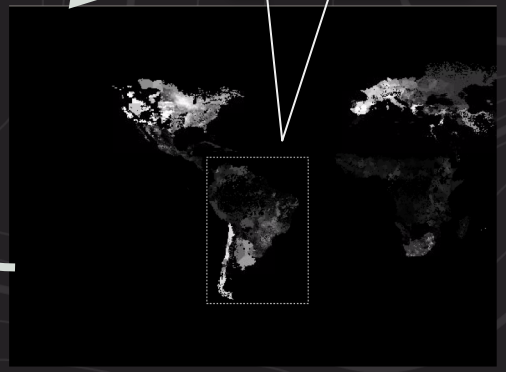
File automatically “loaded” into tool



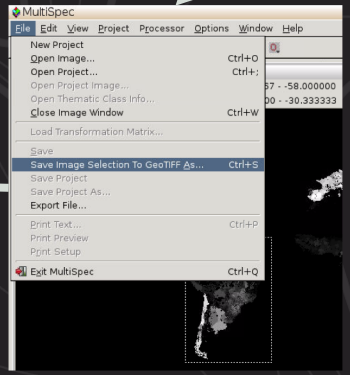
Do arbitrary processing of geospatial files (split into classes in this case)

Processing

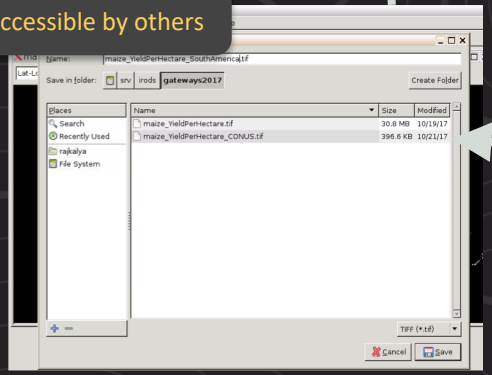
Crop to region of interest

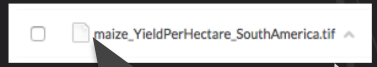


Save result



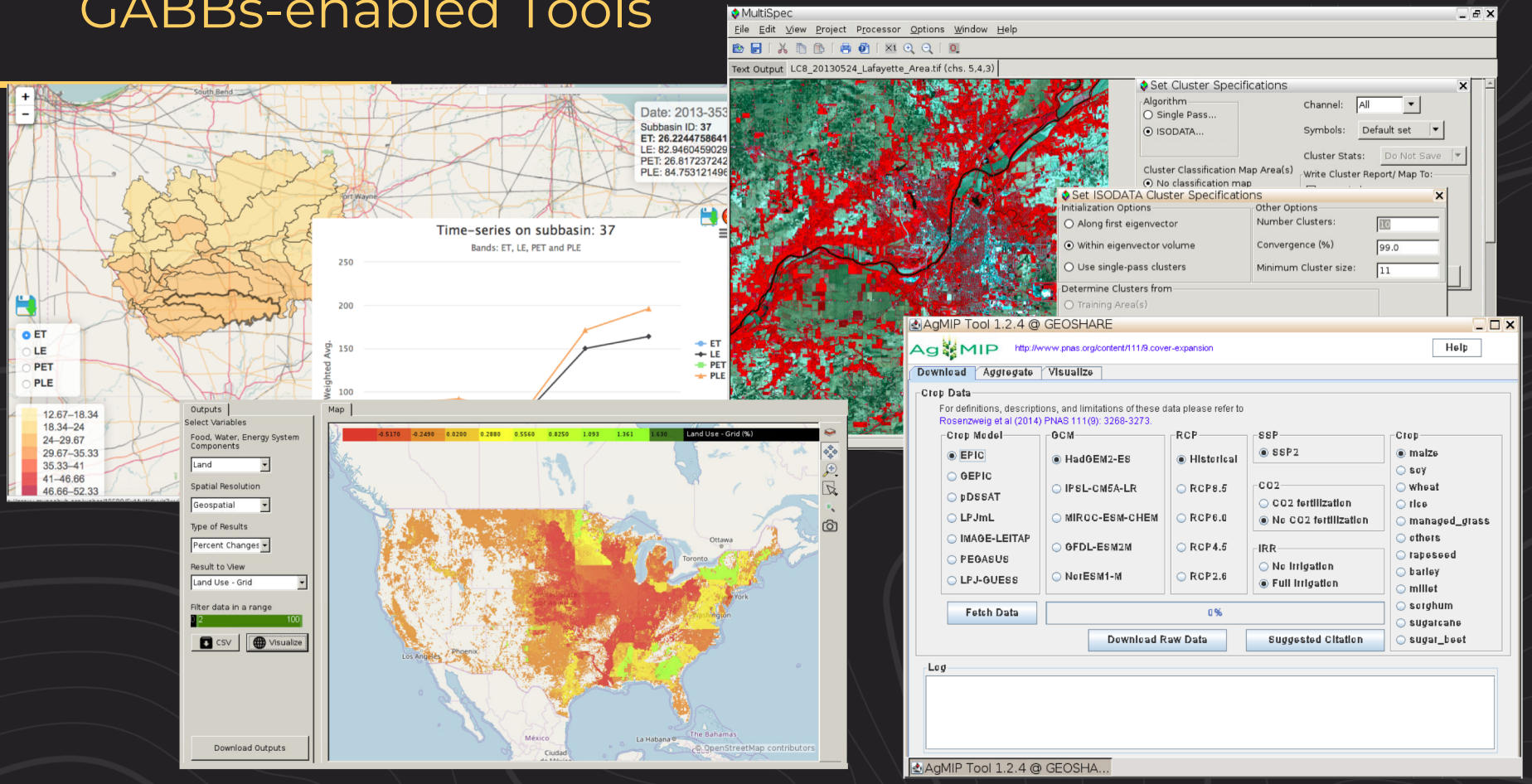
iData filespace visible to any tool





New file now accessible by others

GABBs-enabled Tools



Data Publication (with DOI)

Season-wise irrigated and rainfed crop areas for India around year 2005

By Gang Zhao[1], Stefan Siebert[1]

University of Bonn, Germany

Crop growing area and irrigated fraction for 21 crops in Kharif, Rabi and Zaid seasons for 2005 in 500 m spatial resolution.

Listed in Datasets | publication by group Geoshare

Download Bundle

Additional materials available

Version 1.0 - published on 15 Jan 2015
doi:10.13019/M2CC71 - cite this

Licensed under CC0 Creative Commons

0.0 RANKING

0 citation(s)

Share:

0 review(s) (Review this)

0 questions (Ask a question)

0 wish(es) (Add a new wish)

3398 total view(s), 719 download(s)

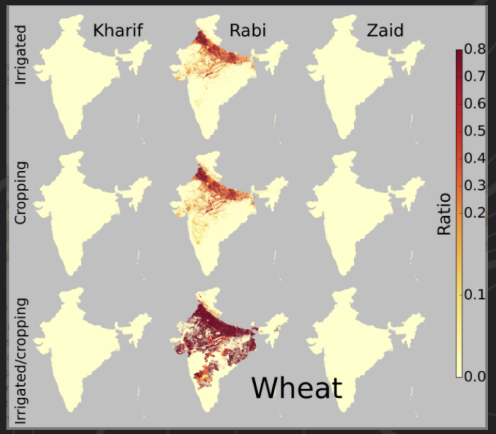
Version **1.0** - published on 15 Jan 2015
doi:10.13019/M2CC71 - cite this

3398 total view(s), 719 download(s)

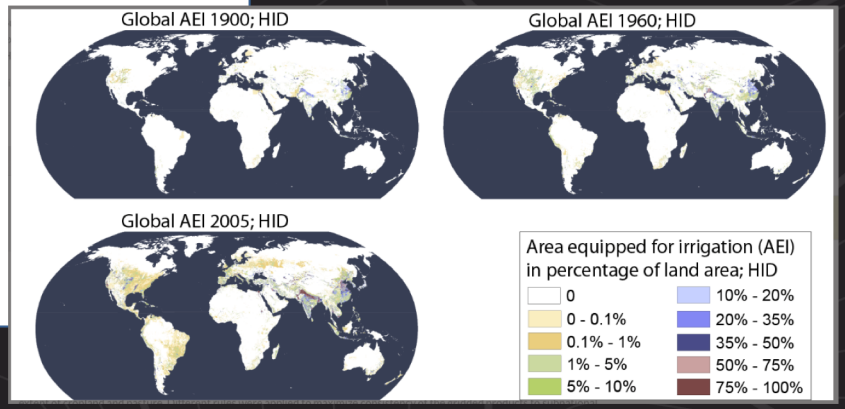
About | Citations | Versions | Supporting Docs | Reviews | Questions | Wishlist | Usage

Description

The spatio-temporal pattern of cultivated crops varies considerably in India and is strongly determined by varying balances of water supply and demand across regions and seasons. The positive balance between precipitation and potential evaporation in the Kharif (monsoon) season lasting from June to October allows to grow one rainfed crop in most parts of the country, while there is a high water deficit in the Rabi season (November to February) and in the pre-monsoon season (Zaid, March to May). Therefore, irrigation is essential for crops in the Rabi and Zaid seasons to reduce drought stress, while it is only supplementary in the Kharif season across most of the country. This kind of knowledge and information are vital for natural resources management and climate change impact assessments. This study aims to map the rainfed and irrigated cropping area for 21 crops by integrating remotely sensed land use and land cover (LULC) data, crop calendars and survey based census data.



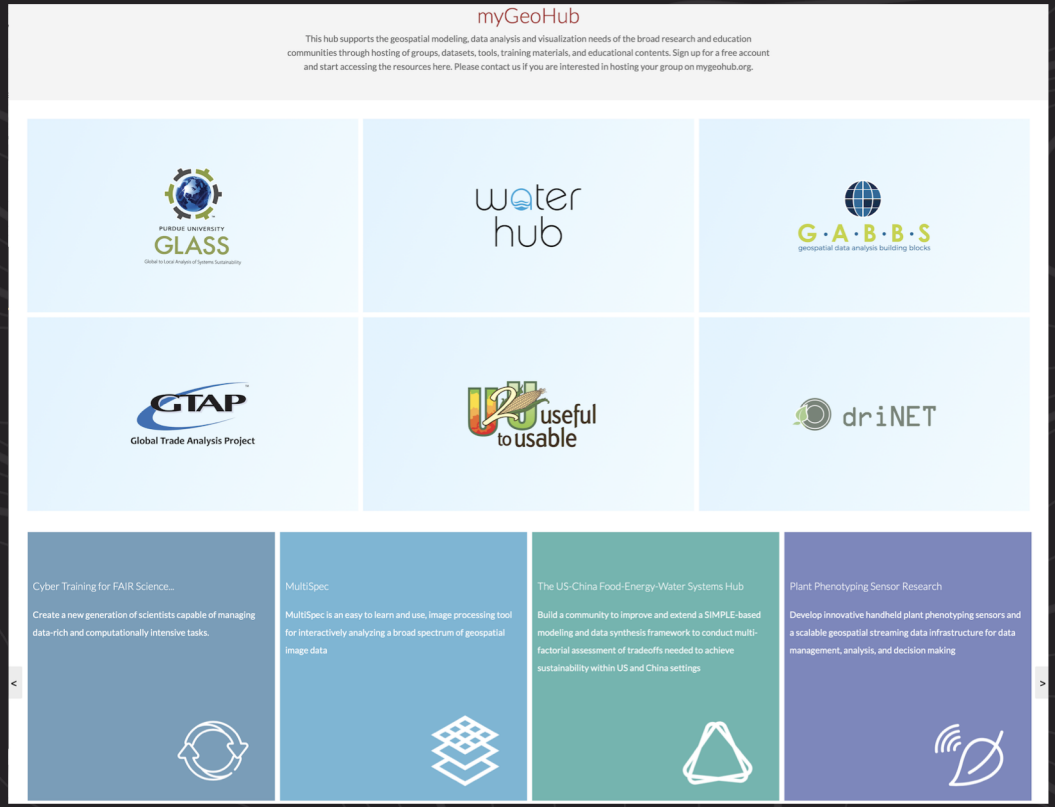
Gridded, spatial-temporal data for India



Another example: Global gridded, spatial-temporal data, 1900-2005

The GeoHub Geospatial Science Gateway

mygeohub.org

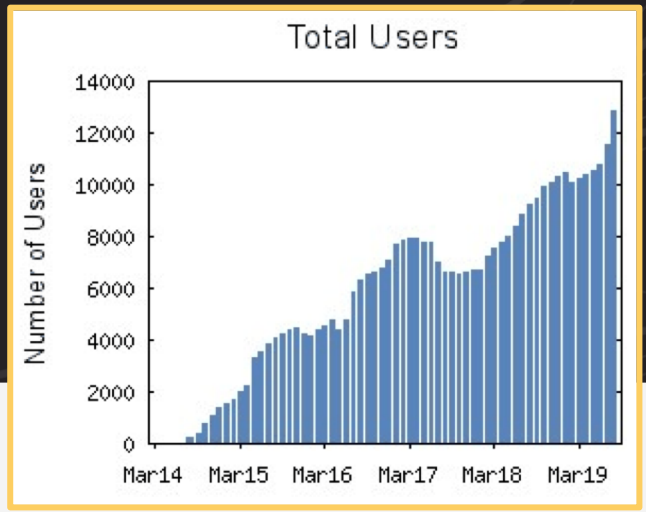


The GeoHub Geospatial Science Gateway

More than 10,000 users each year

Hosting > 10 funded projects

A testbed for new GABBs capabilities



Usage

Usage for prior 12 months

12861
Users

354
Resources

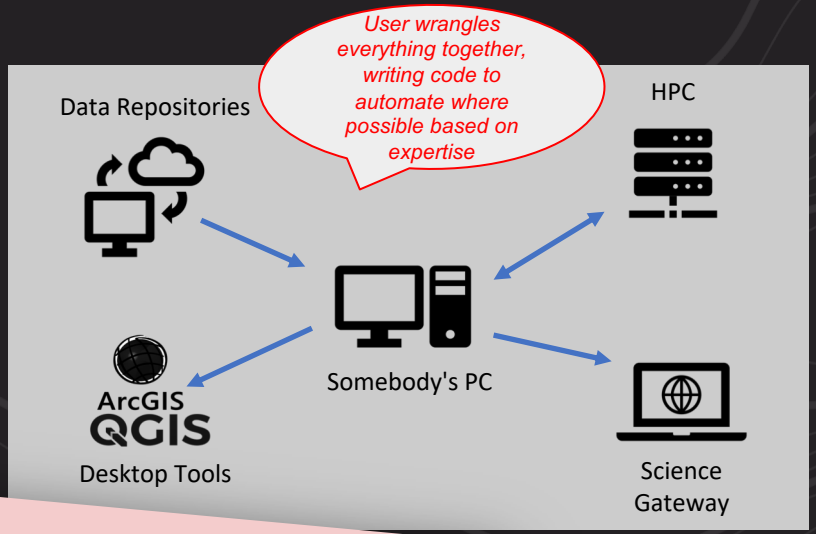
54
Tools

47022
Simulations

Wrangling of data, computation, software, …

OUR DATA WORKFLOW - Ver. ~~1~~ ~~2~~ 3

1. Make sure date is just after 1st or 15th!
2. Go to: ~~usgs.gov~~ prd-tnm.s3.amazonaws.com/index.html?prefix=StagedProducts
3. Browse: Hydrography…NHDPlusHR…Beta…GDB.
4. Download NHDPLUS_H_01##_HU4_GDB.zip where ## is **02 to 14**.
5. Unzip it - WARNING: Have enough space!!!
6. Run our tool. WARNING: Takes a **looooong** time!!! **DO NOT TURN OFF PC!!!**
7. Upload output files to cluster - Note: wait until all successful!
8. Kick off our standard jobs.
9. ~~Occasionally check 'em~~. Wait for email(s)???
10. Download new images.
11. Ask ~~Fred~~ to upload to website. Mary?
12. Tell everybody there's new stuff.



Process your files before running our tool

- get the latest code from ~~Jaewoo~~ Nicole for filtering the input data

Note: that's Windows code - use the lab desktop

- you need to get the maps from the group folder /depot/lyllegroup/project1/maps/exp1/ …. for aggregation

Data Challenges in Being FAIR

Even in the age of large computational resources, research is faced with:

- Manual workflows, or custom-made automated processing
- Coding data acquisition and processing requires diverse software knowledge
- Code requires frequent modification on data provider technology changes
- Cross-domain research requires domain-specific data format knowledge
- Workflows seldom reproducible

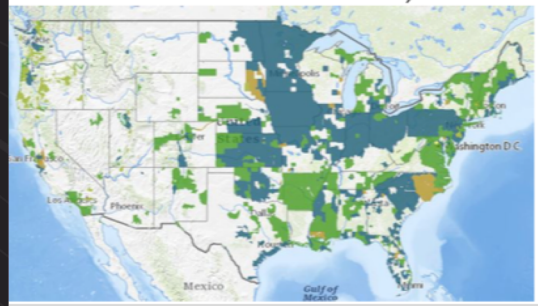
Capturing complex data pipelines (example)

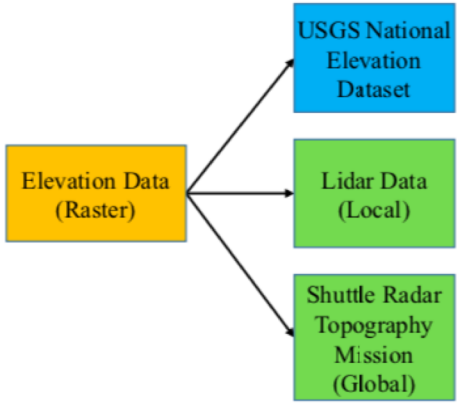
Potential pitfalls with other elevation data sources

- 1/9th arc-second or LiDAR or 1-m DEM not available nationally
- Pre-processed Lidar available in some states (IN, OH, MN, NC)
- Create Lidar acquisition tool where available or not?
- Conversion of Lidar point cloud to bare-earth DEM is an issue
- SRTM → available globally at different resolutions
- Probably need to re-project user shapefile to USGS coordinates first

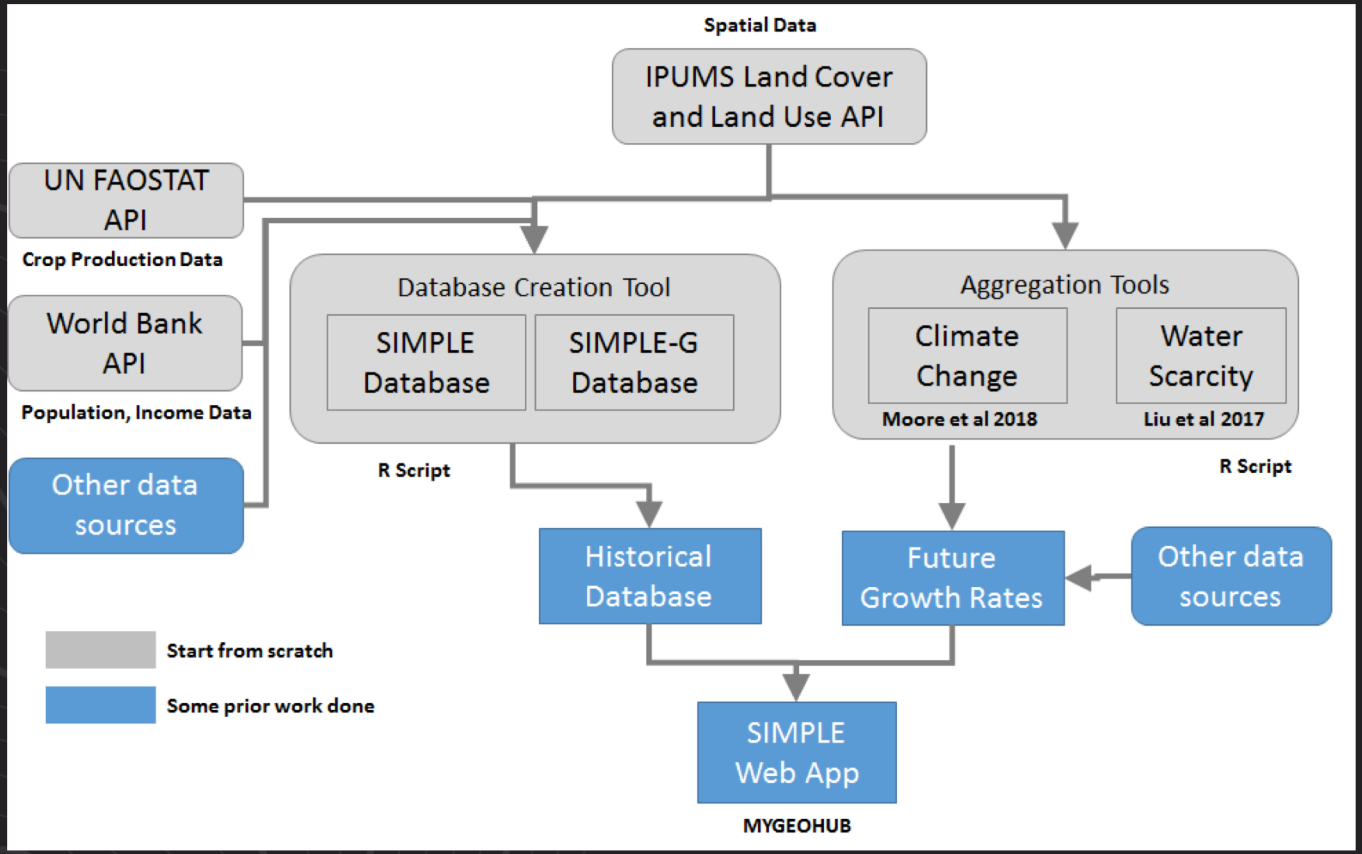
1-m DEM availability

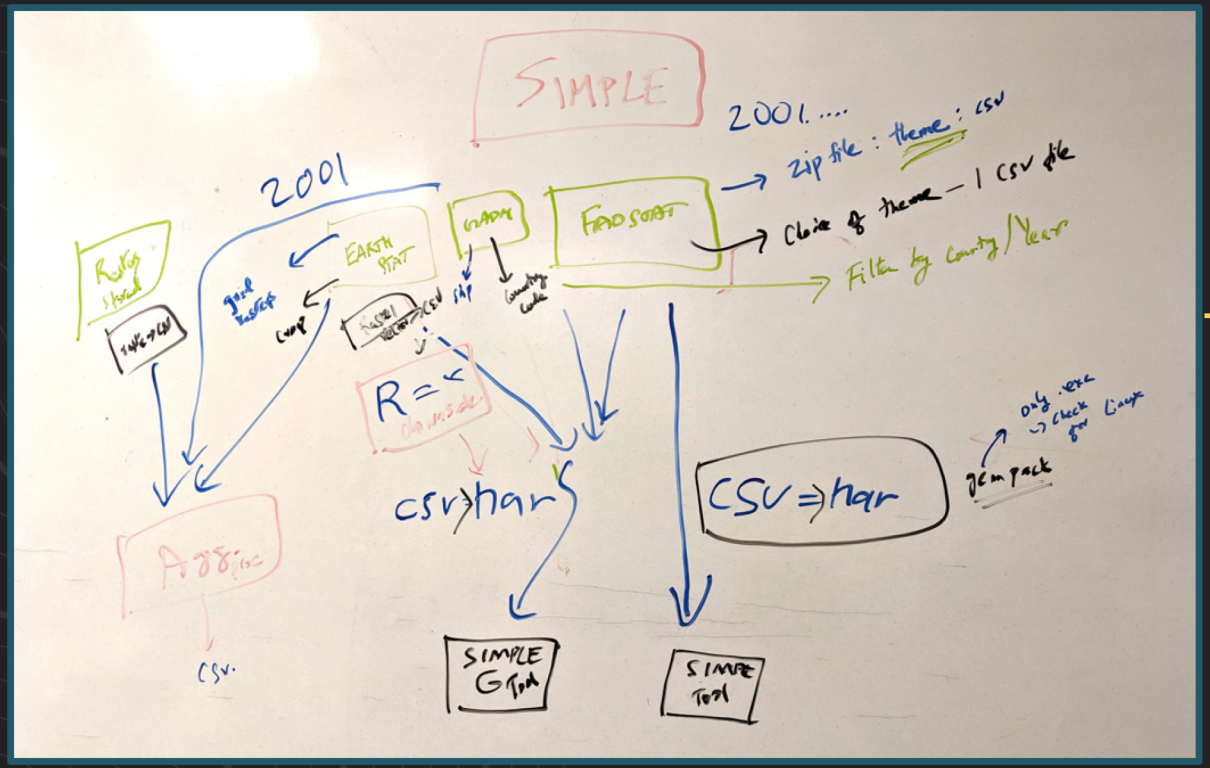
Lidar Point cloud availability





Example Workflow





There are a lot more details

GABBs 2.0: GeoEDF -- Vision

Create an extensible geospatial data framework that will address the challenges by providing seamless connections among platforms, data and tools, hence making valuable, large scientific and social datasets usable directly in scientific models and tools.

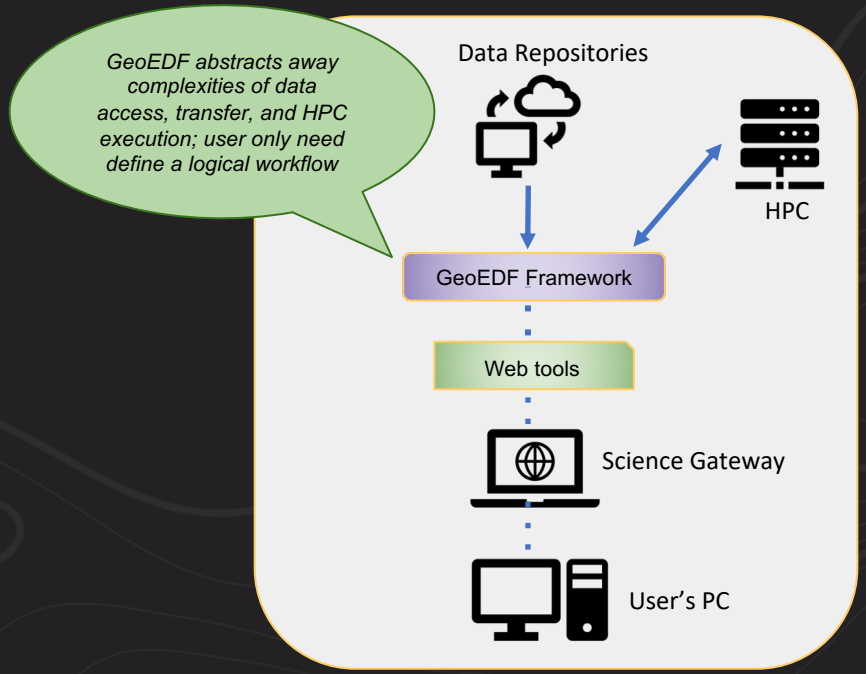
The ultimate goal is to put easy-to-use tools and platforms into the hands of researchers and students to conduct scientific investigations following FAIR science principles.

FAIR = Findable, Accessible, Interoperable, Reusable

Vision: After GeoEDF

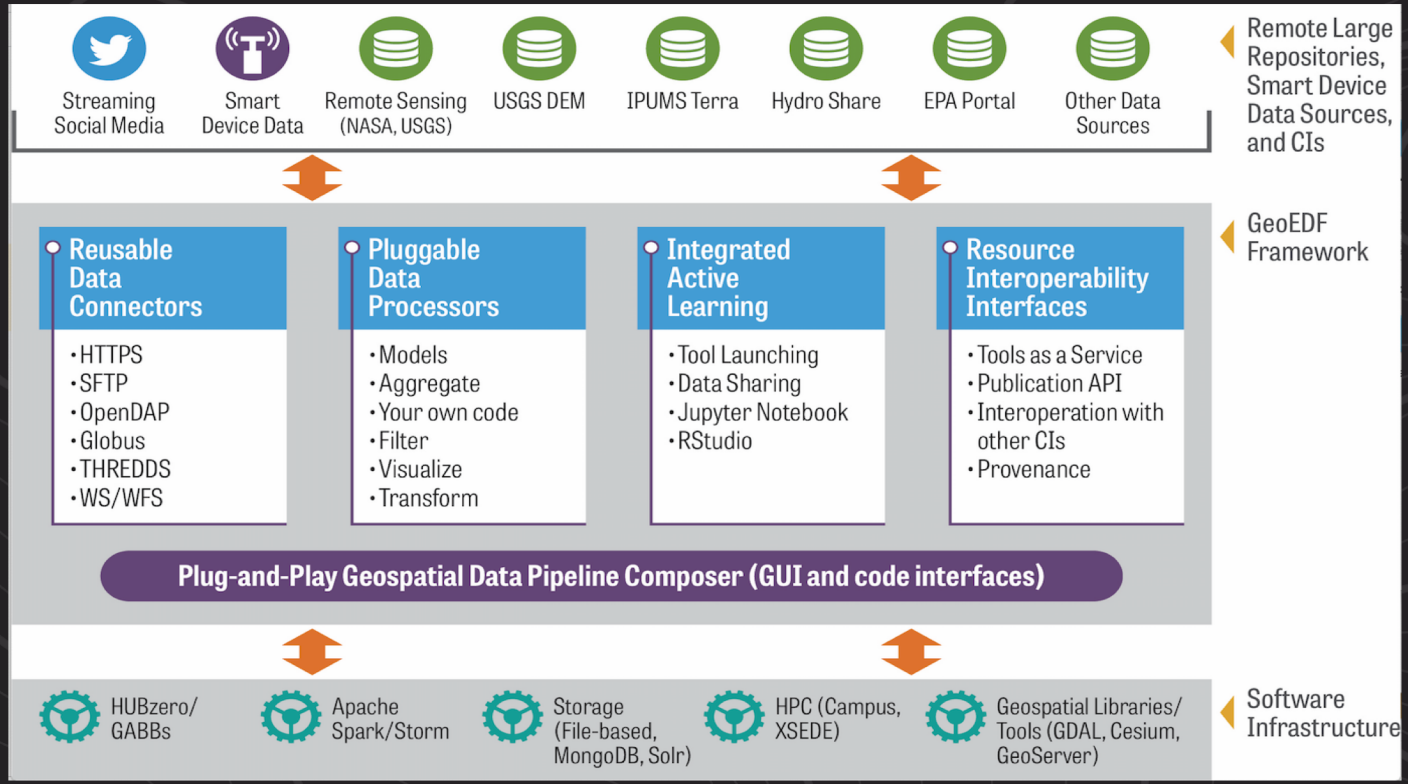
OUR DATA WORKFLOW - Final

1. *Go to the science gateway*
2. *Define "my_workflow.yml" (or use tool GUI if needed)*
3. *Ask GeoEDF to execute!*
4. *Data and workflow automatically published to science gateway*



- Automated, secure, logged process running on dedicated infrastructure - You can log off!
- Leverage building blocks from existing workflows
- Data transfer and HPC execution abstracted away
- Automatic provenance capture and data annotation for future discoverability, reproducibility

GeoEDF High-Level View



Deliverables

CI

Plug-and-play geospatial data framework, open-source packages installable on CI platforms

Science

Scientific workflows composed of reusable building blocks

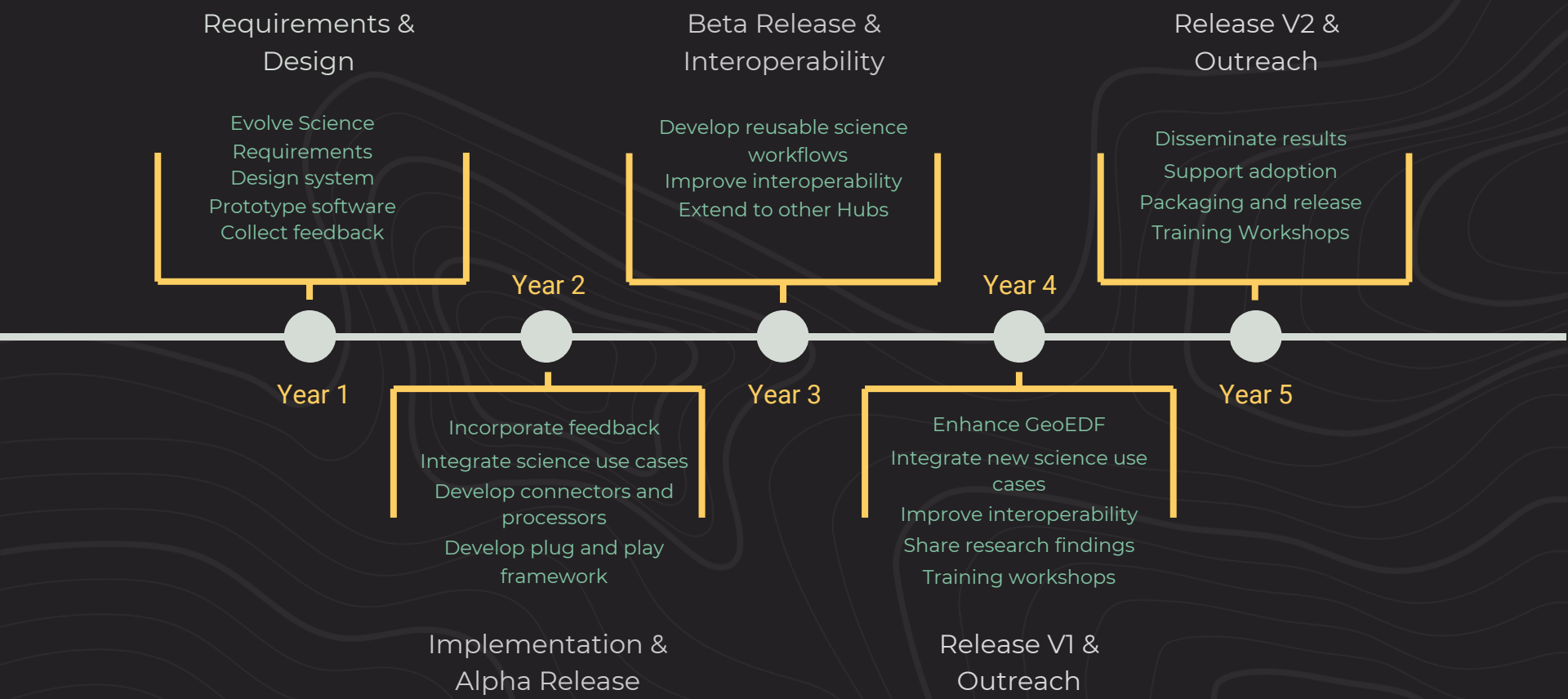
FAIR

Enhanced HUBzero publication, course, tool/data linkage for FAIR science

Outreach

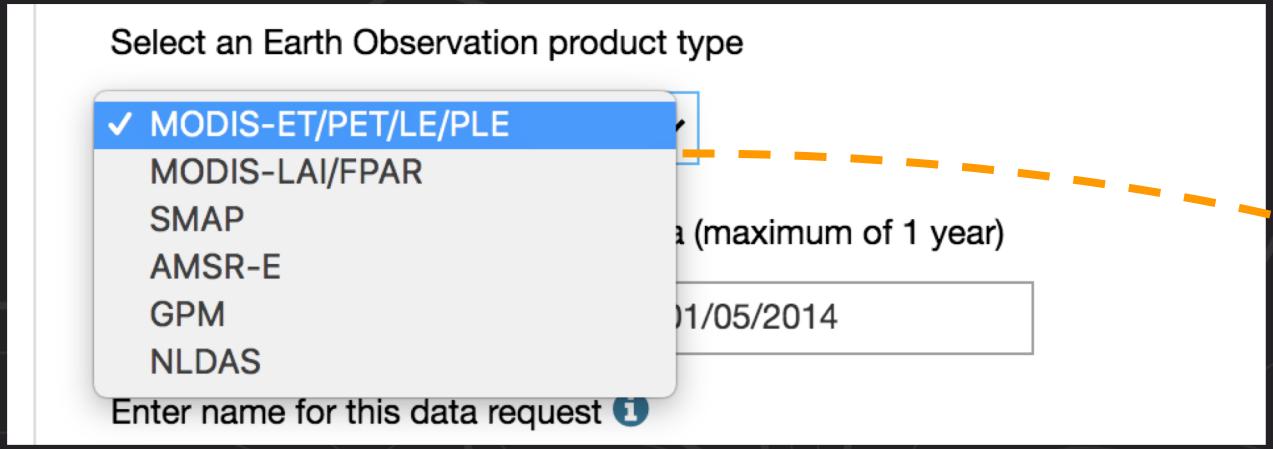
Training materials, CI interoperability, workshops, support

Project Timeline



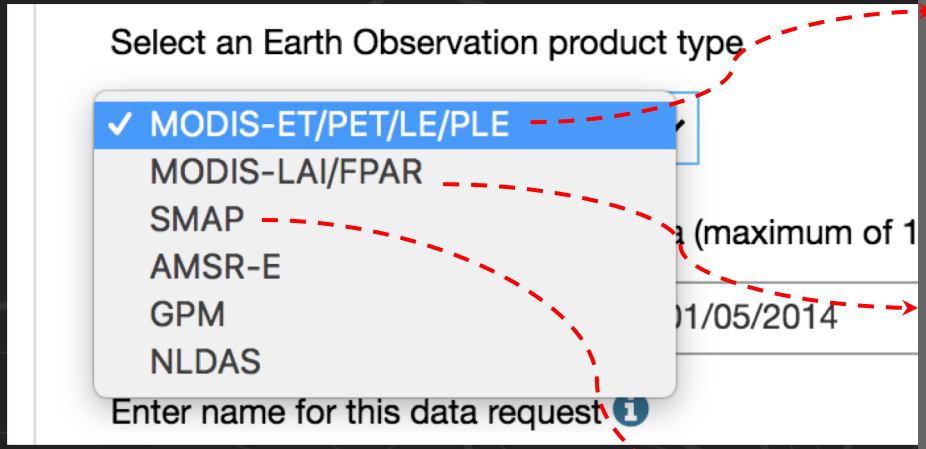
Design and Cyberinfrastructure

Workflow Example I



Mask with shapefile, compute weighted aggregate for each polygon

Workflow Example I - Opportunity

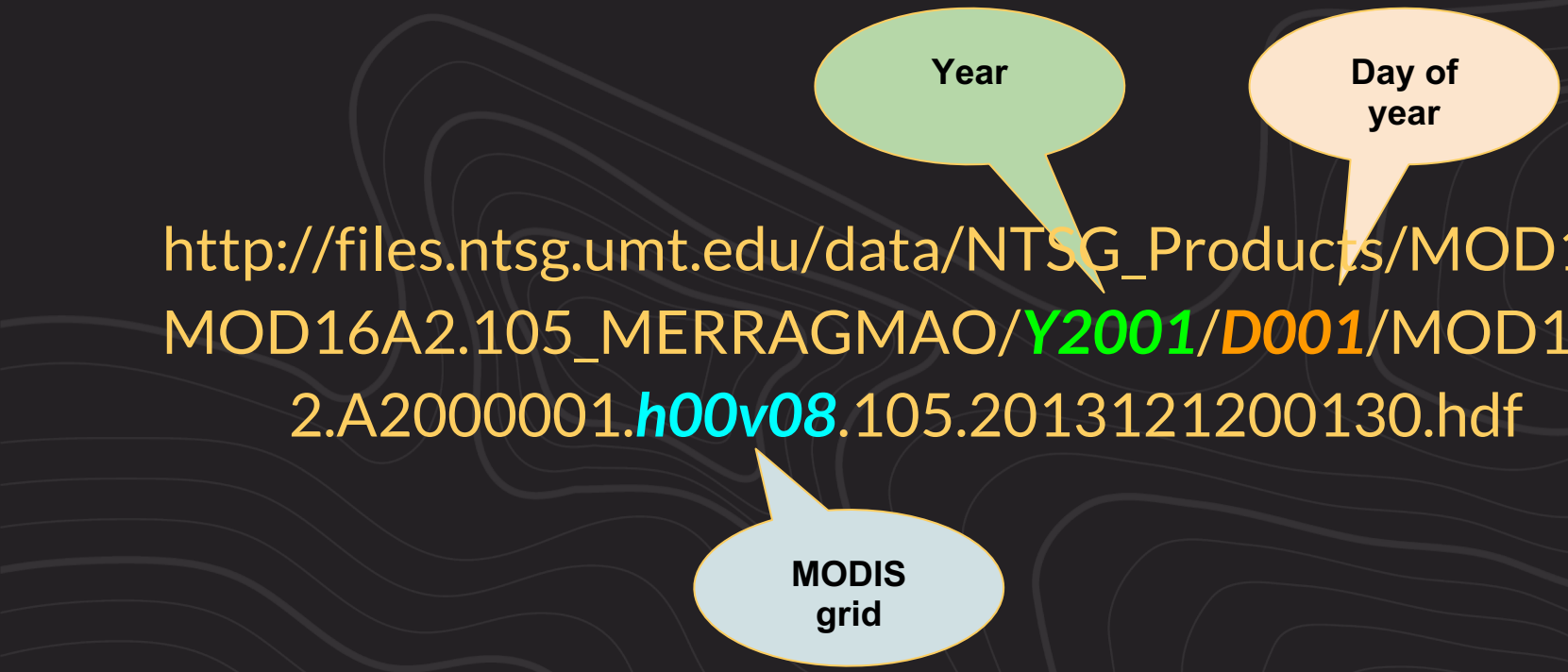


http://files.ntsg.umt.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/Y2001/D001/MOD16A2.A2000001.h00v08.105.2013121200130.hdf

https://e4ftl01.cr.usgs.gov/MOTA/MCD15A3H.006/2002.03.19/MCD15A3H.A2002193.h07v07.006.2015149100709.hdf

https://n5eil01u.ecs.nsidc.org/SMAP/SPL4SMGP.003/2015.03.31/SMAP_L4_SM_gph_20150331T013000_Vv4030_001.h5

Workflow Example I - Opportunity



Workflow Example II

```
def GetNED(NL, WL):
    name1 = "n"+NL+"w"+WL
    address = "ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/Elevation/1/ArcGrid/USGS_NED_1_"
    url_final = address + name1 + "_ArcGrid.zip"
    print(url_final)
work_folder_name = os.path.join(input_folder_name, "WorkFolder")
if os.path.exists(work_folder_name) == False:
    os.mkdir(work_folder_name)
boundary_path = os.path.join(input_folder_name, boundary_file)
input_crs = QgsVectorLayer(boundary_path, '', 'ogr' ).crs().authid()
#processing.run('qgis:reprojectlayer',{'INPUT': full_input_path, 'TARGET_CRS':'EPSG:10267
processing.run('native:reprojectlayer',{'INPUT': boundary_path, 'TARGET_CRS':'EPSG:4326',

input_list = [os.path.join(work_folder_name, cur_raster) for cur_raster in raster_names]
print("Merging Raster...")
processing.run("gdal:merge", {'INPUT':input_list, 'OUTPUT':work_folder_name + "/merged_rast.tif"})
print("Projecting Raster...")
processing.run('gdal:warpreproject', {'INPUT': work_folder_name + "/merged_rast.tif", 'TARGET_CRS':
print("Clipping Raster...")
processing.run('gdal:cliprasterbymasklayer',{'INPUT': work_folder_name + "/proj_rast.tif", 'MASK':
print("DEM prepared successfully!!!")
```

Workflow Example II - Opportunity

Get NED from USGS

```
def GetNED(NL, WL):
    name1 = "n"+NL+"w"+WL
    address = "ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/Elevation/1/ArcGrid/USGS_NED_1_"
    url_final = address + name1 + "_ArcGrid.zip"
    print(url_final)
```

Reproject watershed shapefile

```
work_folder_name = os.path.join(input_folder_name, "WorkFolder")
if os.path.exists(work_folder_name) == False:
    os.mkdir(work_folder_name)
boundary_path = os.path.join(input_folder_name, boundary_file)
input_crs = QgsVectorLayer(boundary_path, '', 'ogr' ).crs().authid()
#processing.run('qgis:reprojectlayer',{'INPUT': full_input_path, 'TARGET_CRS':'EPSG:10267
processing.run('native:reprojectlayer',{'INPUT': boundary_path, 'TARGET_CRS':'EPSG:4326',
```

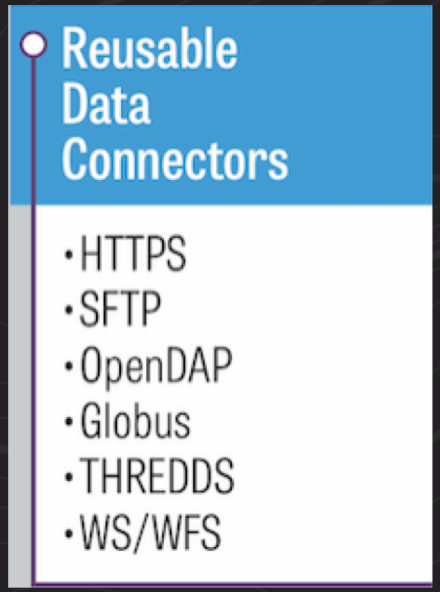
Mosaic -> reproject -> clip raster(s)

```
input_list = [os.path.join(work_folder_name, cur_raster) for cur_raster in raster_
print("Merging Raster...")
processing.run("gdal:merge", {'INPUT':input_list, 'OUTPUT':work_folder_name + "/merged_rast.tif"})
print("Projecting Raster...")
processing.run('gdal:warpreproject', {'INPUT': work_folder_name + "/merged_rast.tif", 'TARGET_CRS':
print("Clipping Raster...")
processing.run('gdal:cliprasterbymasklayer',{'INPUT': work_folder_name + "/proj_rast.tif", 'MASK':
print("DEM prepared successfully!!!")
```

Data Connectors

❖ What are they?

- ➢ Help retrieve remote data (NASA, USGS, field sensors, etc.) and make available in scientific workflows
- ➢ Abstract away specifics of implementation
- ➢ Data sources, sinks in a workflow



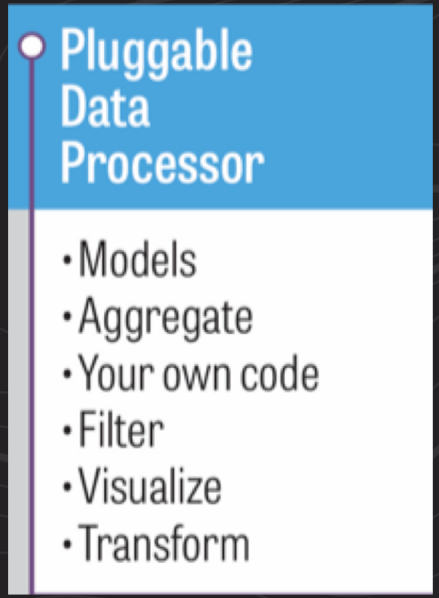
Remote Data Sources (under consideration)

NASA	MODIS, SMAP, other Earthdata DAACs
USGS	Elevation, land use, hydrography, Gage, NLDI
USDA	Soil, land cover, land use
CUASHI	Rainfall, Hydroshare resources
EarthStat	Crop data
FAO	Arable land, harvest data
CIESIN	Population data
EPA	Water quality
Others (no API yet)	Open Data Cubes, Google Earth Engine, ESS-Dive

Data Processors

❖ What are they?

- ➢ Data transformers that can be plugged into a workflow
- ➢ Range from simple geospatial data transformation to scientific simulation models
- ➢ Pre and post processing



Processing Operations (under consideration)

Domain Independent	Reproject, resample, format transformation, filter, mosaic, clip/mask, aggregate (spatial & temporal), visualization, reclassification
Hydrology	Terrain analysis, flood models
Digital Ag	Query, spatial/temporal filter, ML training, decision support
Sustainability	Downsample, (weighted) aggregate, FEWS models

Plug-and-play Workflow Composer

Plug-and-Play Geospatial Data Pipeline Composer (GUI and code interfaces)

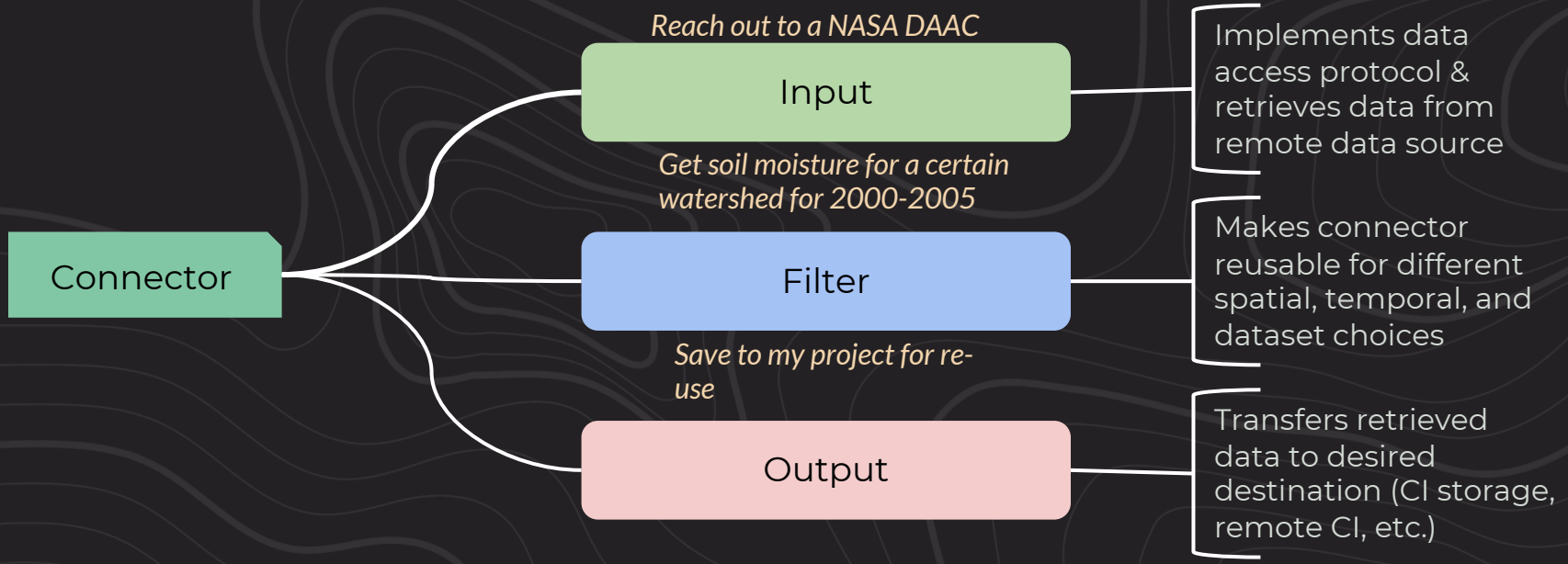
❖ What is this?

- ➢ Framework for composing data connectors and processors into scientific workflows
- ➢ Transforms abstract workflow into actual workflow executing on heterogeneous compute

01

Implementation

Data Connectors - Design



Data Connectors - Example Definition

```
Input:
   NASAInput:
         url:

http://files.ntsg.umt.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/%{file}
         user: rkalyana
         password:
Filter:
   file:
       PathFilter:
           pattern: 'Y%{year}/D001/*.h00v08*.hdf'
   year:
       DateTimeFilter:
            pattern: '%Y'
            start: 01/01/2000
            end: 12/31/2005
            period: 1Y
```

Python Input class

Variable to be bound by filter

Filter returns string bindings for variable

Filter params can also have variables

Data Processors - Example Definition

Python processor class implementing masking operation

```
HDFShapefileEOSMask:
        hdffile: /data/workflow263/mod16Y2001D1T1200.h00v08.hdf
        shapefile: /home/rkalyana/subs1.shp
```

Processor specific params; validated during instantiation

GeoEDF Workflow - Example Definition

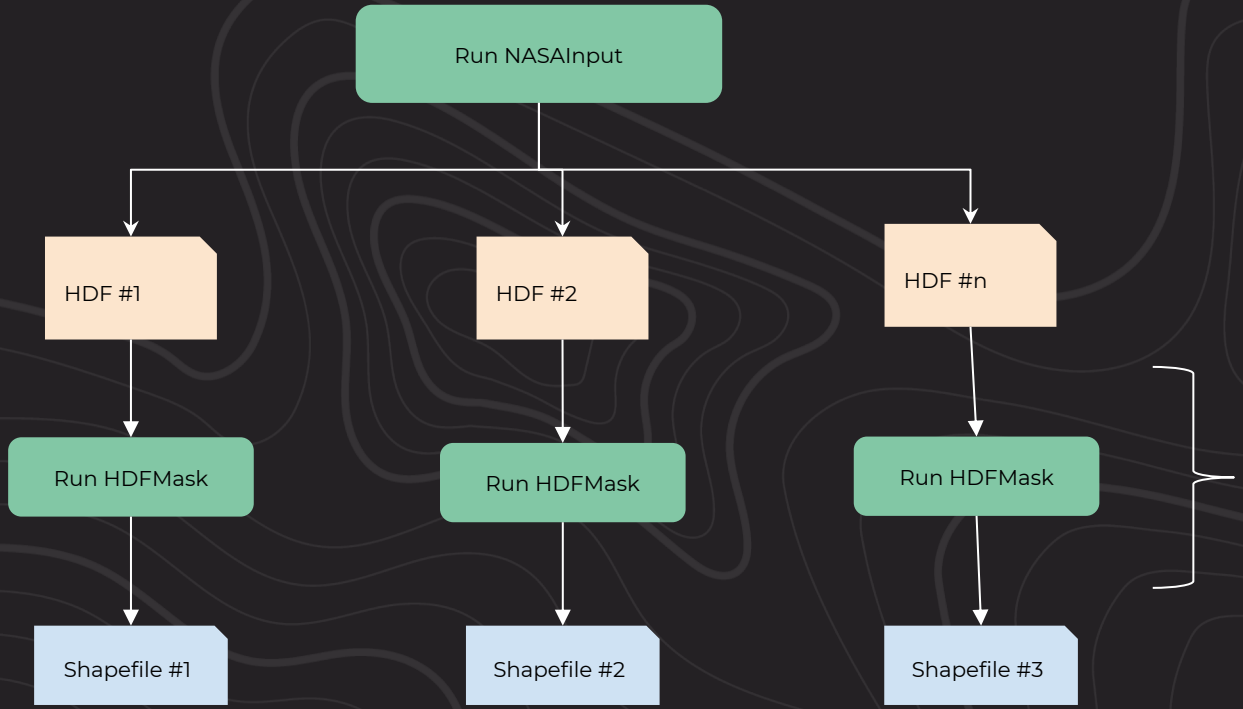
```
$1:
   Input:
      NASAInput:
            url:
               http://files.ntsg.umt.edu/data/NTSG_Products/MOD16/MOD16A2.105_MERRAGMAO/%{file}
            user: rkalyana
            password:
   Filter:
      file:
          PathFilter:
              pattern: 'Y%{year}/D001/*.h00v08*.hdf'
      year:
          DateTimeFilter:
               pattern: '%Y'
               start: 01/01/2000
               end: 12/31/2005
               period: 1Y
 $2:
   HDFShapefileEOSMask:
         hdffile: $1
         shapefile: /home/rkalyana/subs1.shp
```

Workflow stage

Reference output of prior stage

- ★ NASAInput can be used to access any Earthdata-associated repository
- ★ HDFShapefileEOSMask can be applied to any HDF4 or HDF5 file

Actual Scientific Workflow



Parallel compute jobs generated dynamically based on number of data files at each level

Putting It All Together

Workflow Definitions

Users pick and choose different connector & processor classes to define a workflow

(as YAML file/via GUI/through API)

Workflow Execution

Workflow engine transforms declarative specification into concrete Pegasus scientific workflow and executes on heterogeneous compute

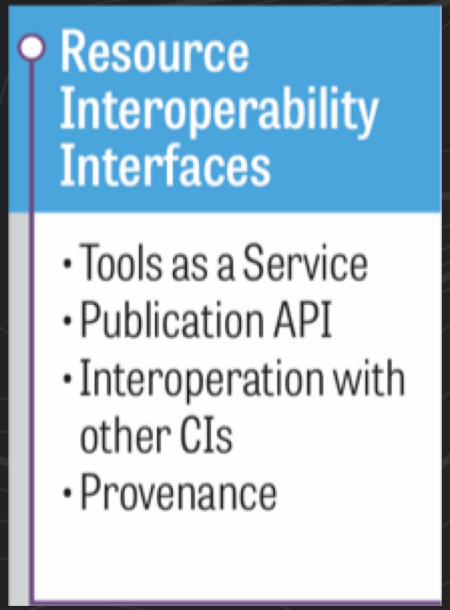
GeoEDF Building Blocks

Users contribute various connector (Input, Filter, Output), and processor classes

Interoperability & FAIR Science

❖ How do we do FAIR?

- ➢ Data publications can be searched using their content metadata, accessed via APIs & used in workflows
- ➢ Automatically track metadata, provenance in workflows
- ➢ Launch tools, workflows seamlessly from a remote CI (with remote data inputs)



Collaboration and Interoperability

Broader Impacts - Community

Usable by especially interdisciplinary research domains

a. Critically important to research supporting the SDGs
b. Data synthesis, multi-scale analysis
c. Lower technology barrier (e.g., seamlessness, extensibility)

Help domain researchers meet new FAIR data/software requirements (e.g., journal, funding agencies)

Help domain science with broader dissemination (e.g., decision makers, public)

Future workforce training (learn and start with good practices)

Opportunities

Interoperate with other CIs

Users contribute connectors and processors

Engage data producers to expose data through connectors

Questions

Questions?

Data to Action Webinar: Increasing the Use and Value of Earth Science Data and Information

October 25th, 2019 | 1:00 pm ET

Putting Data to Work: Building Public-Private Partnerships to Increase Resilience & Enhance the Socioeconomic Value of Data

2020esipwintermeeting.sched.com

2020 Winter Meeting

January 7-9, 2020

Bethesda North Marriott, Bethesda, MD

Engagement Ops.

DISCOVER

Find people and tools to make your data findable, accessible, interoperable, and reusable.

COLLABORATE

Join-in or create a new collaboration area around your Earth science data challenges.

INNOVATE

Utilize small-grant funding to build or expand Earth data technologies.

NETWORK

Extend your network. Build connections across federal agencies, the private setor, and academia.

JOIN

Encourage your organization to join ESIP's 110+ member organizations. Unlock membership benefits: start new collaborations, apply for funding, and more.

Stay up-to-date on all things ESIP by signing up to receive Monday Updates: http://eepurl.com/rJQYn.

https://www.esipfed.org/onepager

Thank you!

Upcoming Webinars

- November 13th at 4 pm ET:
  - Data for our Planet: Increasing the Use and Value of Global Information Infrastructures to Support Resilient Cities, Disaster Risk Reduction and Infectious Diseases (Lesley Wyborn & Erin Robinson)
- Check the webinar homepage: https://www.esipfed.org/webinars.
- Webinar recordings are shared on the ESIP YouTube Channel.