

How to **expand** the **Galaxy** from genes to **Earth** in six simple steps (and live happy)

Raffaele Montella^{1,2}, Alison Brizius², Joshua Elliott², David Kelly², Ravi Madduri^{2,3}, Ketan Maheshwari³, Cheryl Porter⁴, Peter Vilter², Michael Wilde², Wei Xiong⁴, Meng Zhang⁴ and Ian Foster^{2,3,5}

¹*Department of Science and Technologies, University of Naples Parthenope, Naples, ITALY;*

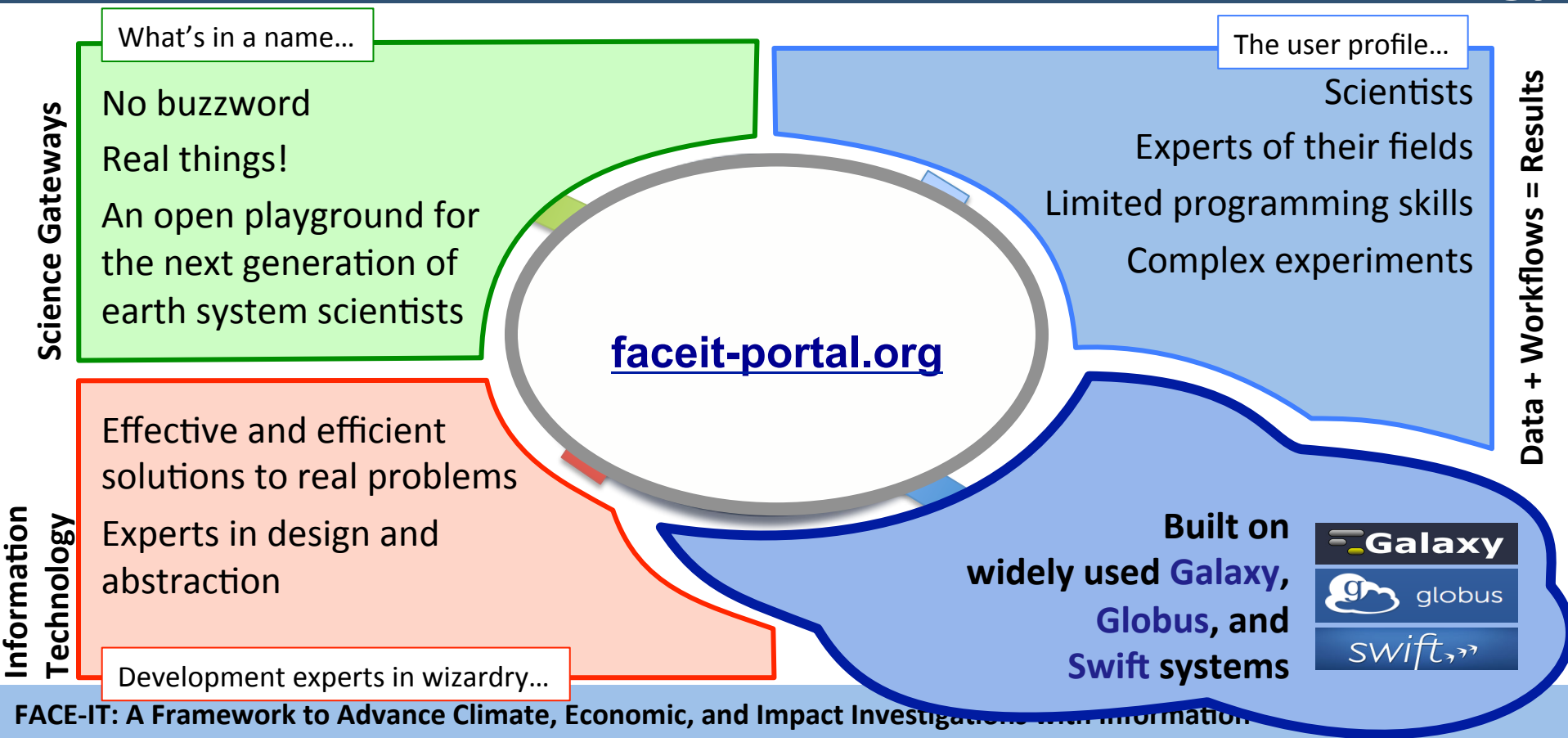
²*Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, Illinois, USA;*

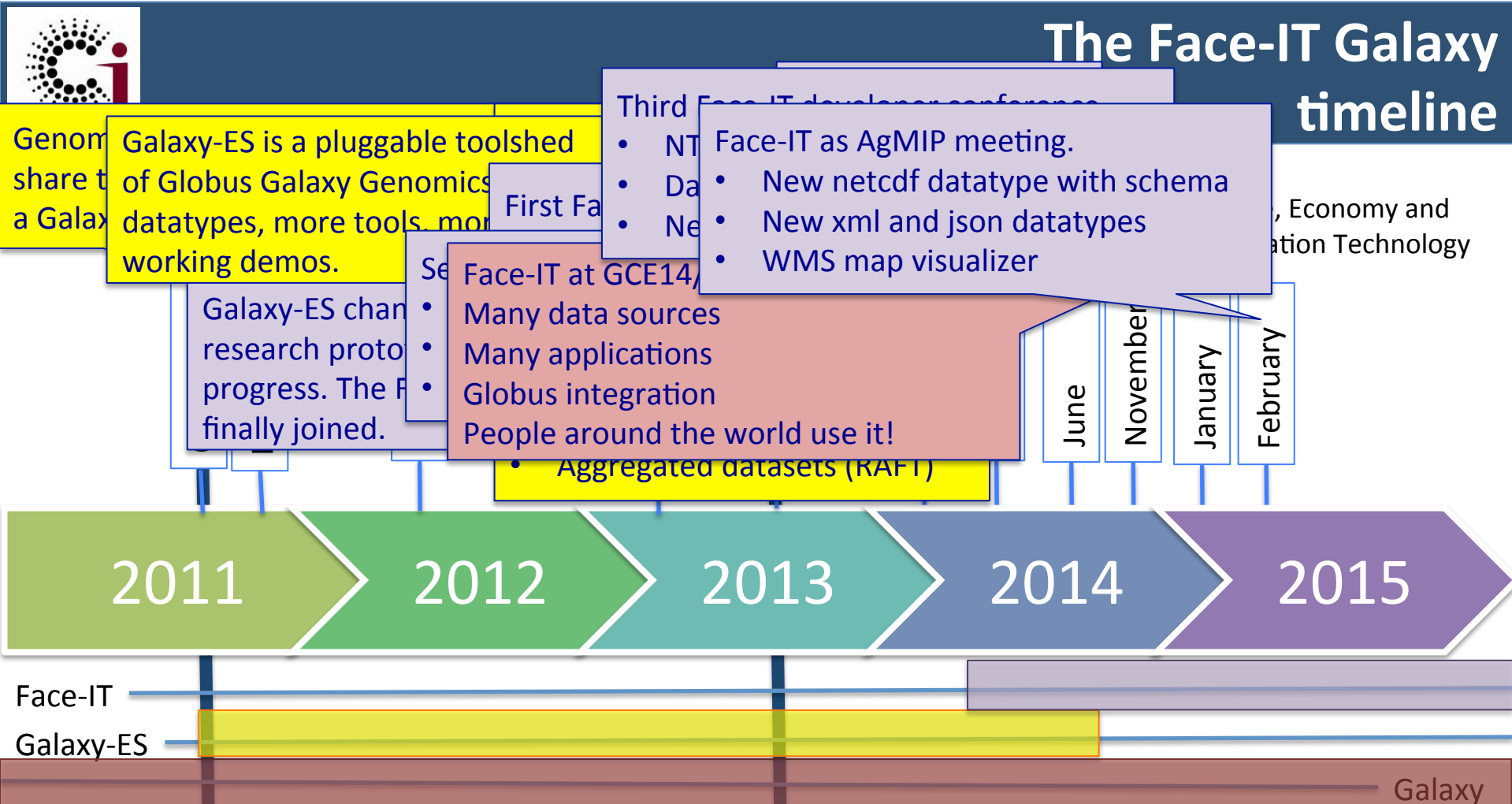
³*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA;*

⁴*University of Florida, Department of Agricultural and Biological Engineering, Gainesville, Florida, USA;*

⁵*Department of Computer Science, University of Chicago, Chicago, Illinois, USA;*

Facing real problems with Information Technology





The Hitchhiker's [Data Analysis] Guide to the Galaxy



Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 3.3 GB

Tools

search tools

Get Data

Easy-SIM

RIA

RIA_Others

Workflows

- All workflows

FACE-IT

Framework to Advance Climate, Economic, and Impact Investigations with Information Technology

FACE-IT is supported by the [NSF cyberSEES](#) program award No.1331782

The Galaxy project is supported in part by [NSF](#), [NHGRI](#), and the [Huck Institutes of the Life Sciences](#).

The Galaxy-ES (Earth System) toolshed is part of the [FACE-IT](#) project.

The FACE-IT framework is being developed out of a collaboration between the University of Chicago Computation Institute's center for Robust Decision-making in Climate and Energy Policy ([RDCEP](#)) and the University of Florida ([ABE/UFL](#)) to meet the needs of several international communities of researchers working on issues around climate change vulnerabilities, impacts, adaptations, and mitigation.

History

Unnamed history

70.7 MB

Dataset

2.0105551-0 latlon.nc4

35

format: gcmLatlon.nc

uploaded GCMlatlon file

History

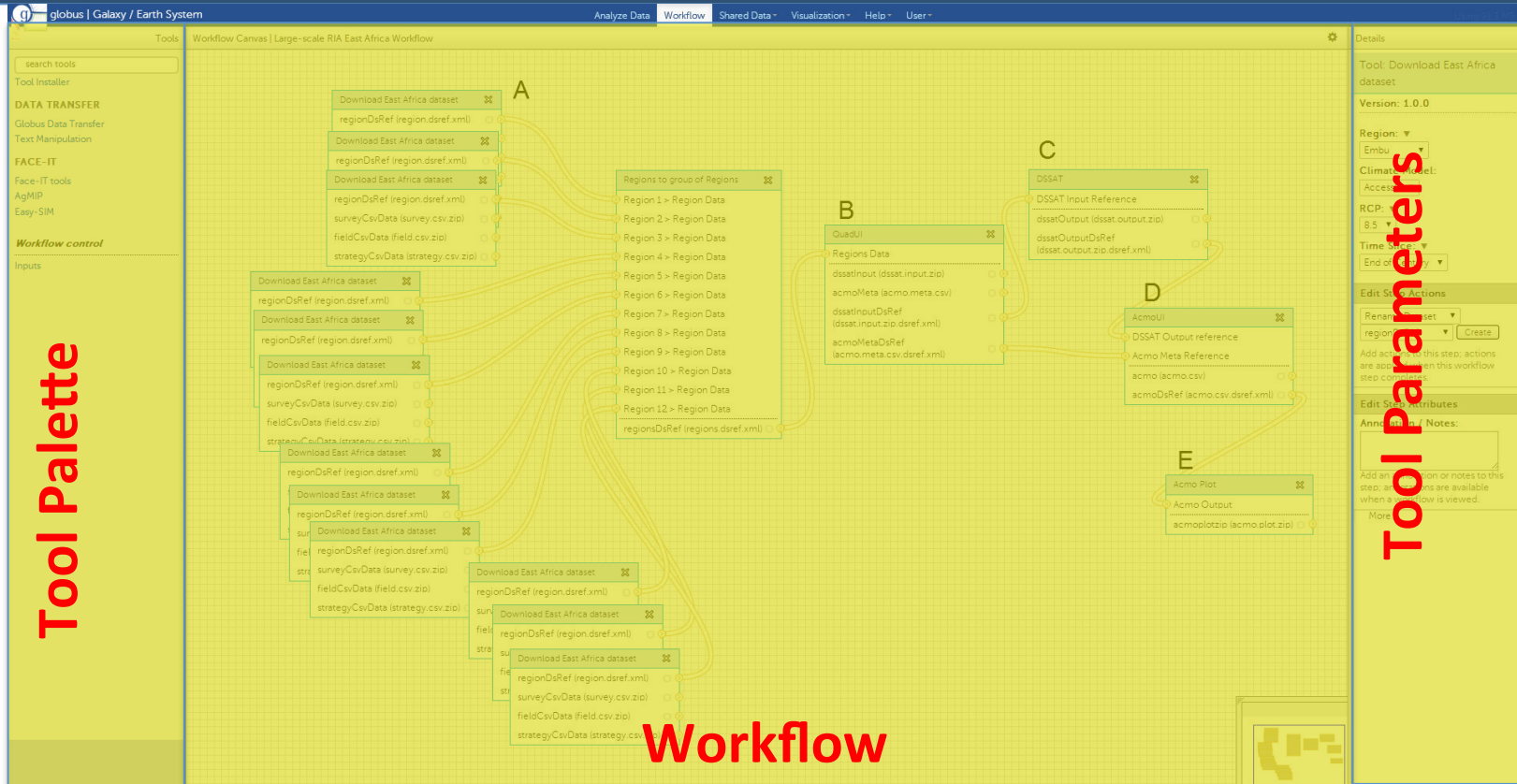
id	name
145	lat
192	lon
12	month
3	decade_hist
9	decade_rcp

Peek Area

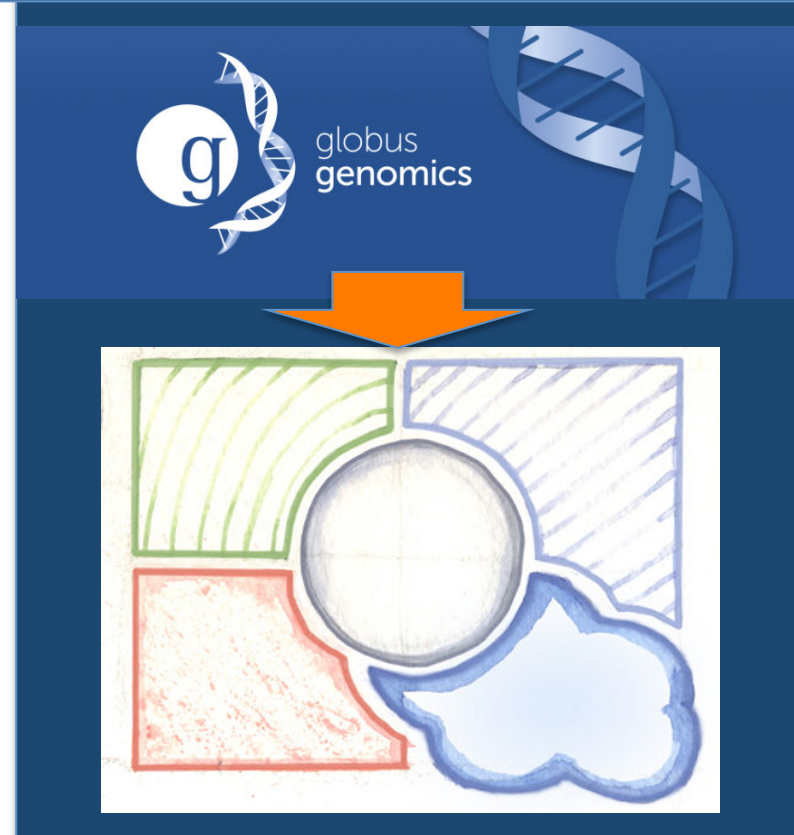
id	name	long_n
145	lat	latitude
192	lon	longitude
12	month	month of
3	decade_hist	
9	decade_rcp	
1.e+20	fwetpr1_hist	
1.e+20	fwetpr1_rcp45	
1.e+20	fwetpr1_rcp85	
1.e+20	meanpr_rcp85	
1.e+20	meantosmax_hist	
1.e+20	meantosmax_rcp45	
1.e+20	meantosmax_rcp85	
1.e+20	meantosmin_hist	
1.e+20	meantosmin_rcp45	
1.e+20	meantosmin_rcp85	

Canvas

The Hitchhiker's [Workflow] Guide to the Galaxy

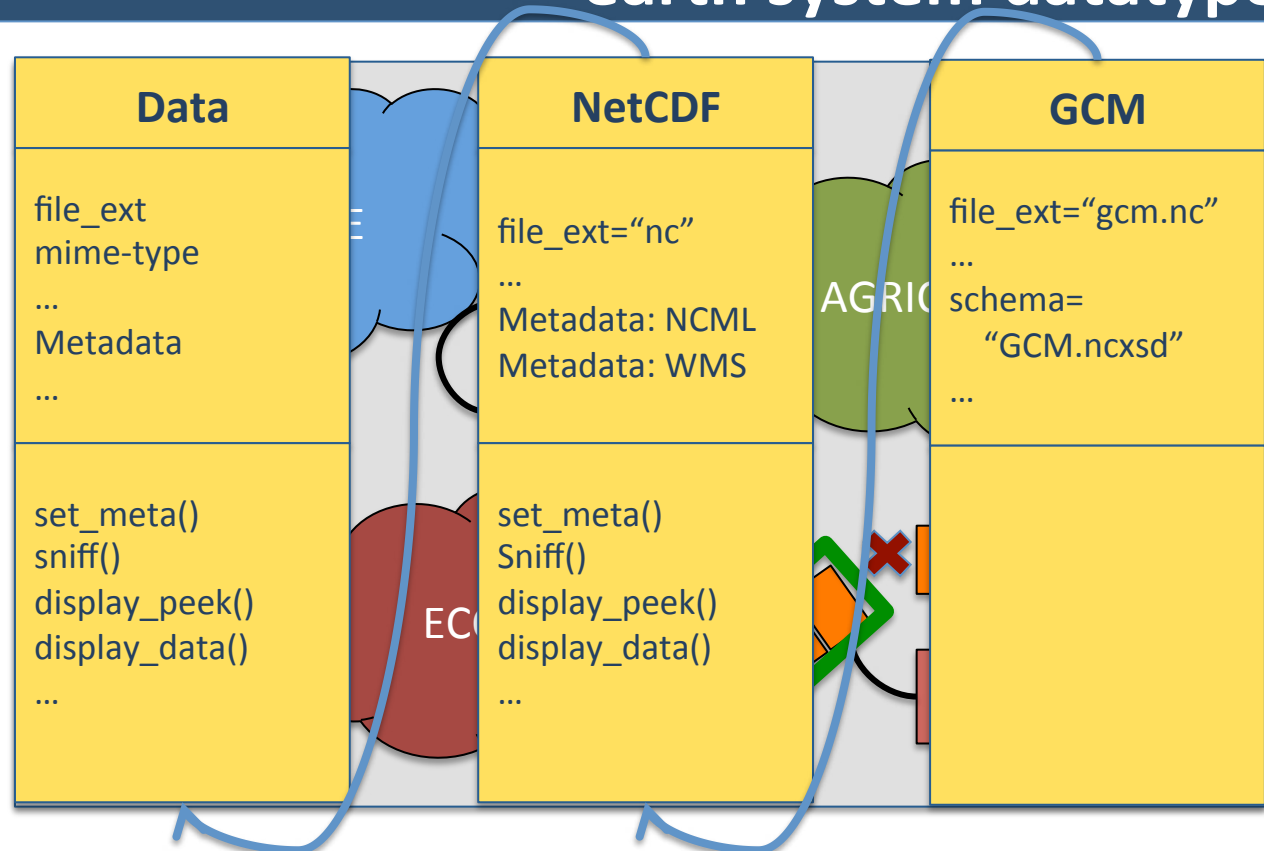


- Datatypes
- Tools
- Tool parameters
- Aggregated datatypes
- Data providers
- Visualizers



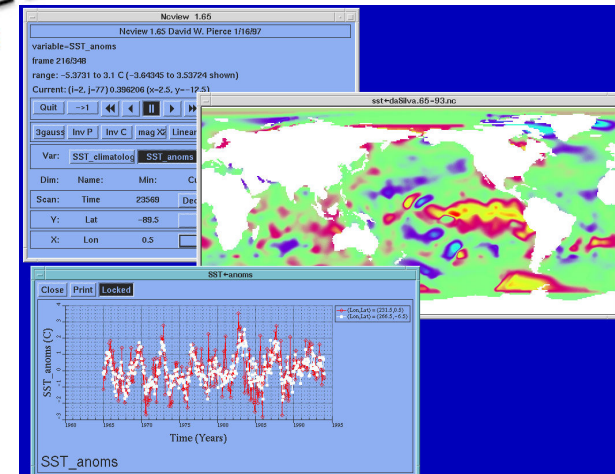
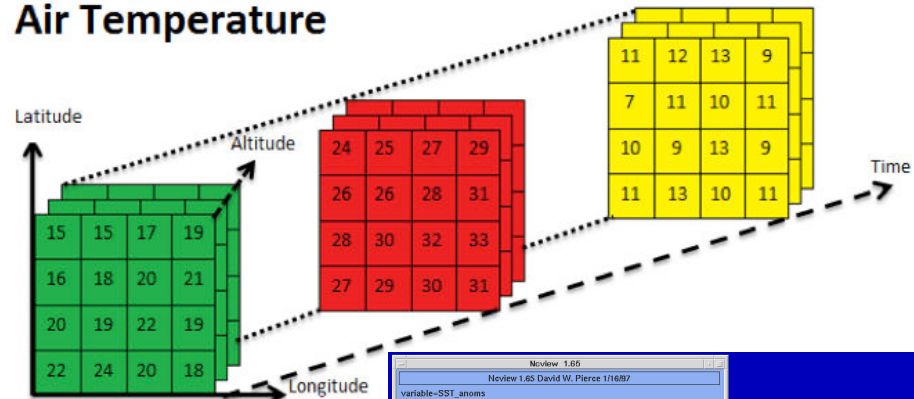
Step ONE: earth system datatypes

- **Datatype:**
the kind of data we want to deal with
- **Dataset:**
the actual data we manage as belonging to a datatype
- If you are thinking about classes and instances in the OOP model you are right!
- Implemented as Python classes



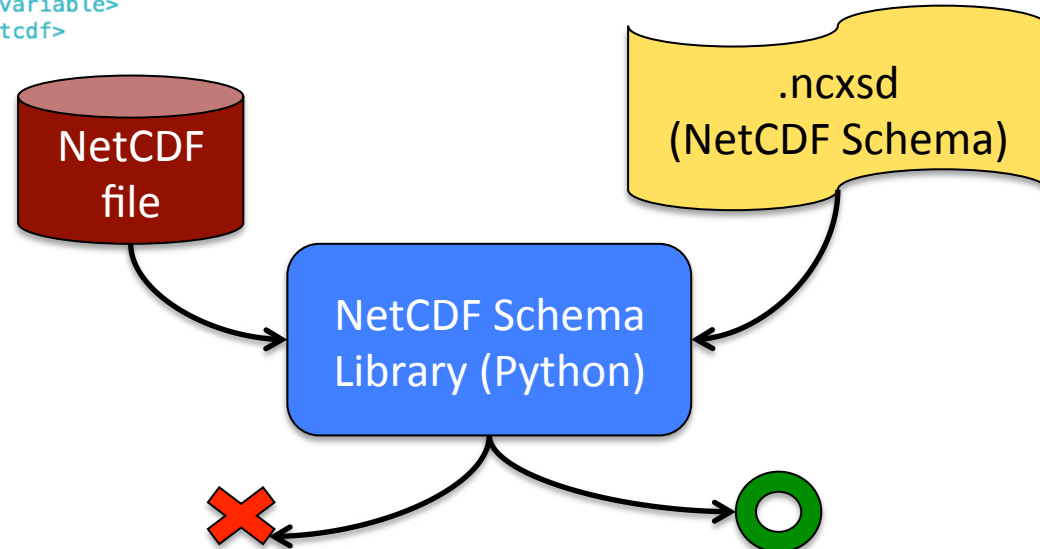
- **NetCDF:**
wide-spread file format for multidimensional environmental data
- Supports unstructured, regular and curvilinear grids
- Dimensions, variables and attributes
- Self descriptive
- Conventions
- Huge amount of data sources, libraries and tools

Air Temperature



- **NetCDF Schema:**
a brand new way to compare and match different NetCDF files.
- Based on wide spread and stable technologies
 - XML Schema
 - NetCDF Markup Language
 - Regular expressions
- Originally built for NetCDF sniffing in Face-IT Galaxy could be something promising...

```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2">
  <dimension name="lat" length="*" />
  <dimension name="lon" length="*" />
  <variable name="meantasmx_rcp45" shape="decade_rcp month lon lat" type="float">
    <attribute name="_FillValue" type="float" value="1.e+20" />
  </variable>
  <variable name="meantasmx_rcp85" shape="decade_rcp month lon lat" type="float">
    <attribute name="_FillValue" type="float" value="1.e+20" />
  </variable>
</netcdf>
```



- Tool:
Is a computing **process** fed by one or more datasets **producing** one or more **datasets**
- It is wrapped over **any** kind of **executable**
- Running by naïve local **scheduler**, **super-computers**, virtual machines somewhere in the **cloud**.
- Each input and output is data **typed**
- It is **defined** using XML

FACE-IT | Galaxy

Tools

Get Data

Climate

Models

Easy-SIM

Experiment Creator Create a PSIMS experiment. The experiment file allows you to control the number of years in a simulation, the crop, irrigation management, and more

Plot Output Generate a plot from the NetCDF output of an Easy-SIM workflow

Model Translator Convert experiment, weather, and soil data from PSIMS format to the native formats used by each crop model

Model Output to NetCDF Convert crop model output to PSIMS NetCDF

RIA tools

AgMIP

QuadUI (version 1.3.1)

Input type:
csv_zip

Survey CSV Data:
☐

Supplements
Add new Supplement

DOME type:
Seasonal Strategy

Field Overlay DOME:
☐

Seasonal Strategy DOME:
☐

Linkage between field and DOME:
☐

Cultivar Source:
Customized
No data? See tip below

Execute

A tool in data analysis

The same tool
in a workflow

QuadUI

- Survey CSV Data
- Supplement 1 > Supplementary Data
- Field Overlay DOME
- Linkage between field and DOME
- Combined_ACE_PLUS_DOME (json)
- cultivar_file_package (zip)

The tools palette

[changing the order of running dimensions] new tools

- The tool executable is run in a scratch directory
- By default input and output datasets are managed “in place”
- Data-typing is strictly enforced

```
<variable name="fwetpr1_rcp45"
  shape="decade_rcp month lon lat"
  type="float">
```

...

```
</variable>
```

GCMlatlon

GCM

```
<variable name="fwetpr1_rcp45"
  shape="decade_rcp month lat lon"
  type="float">
```

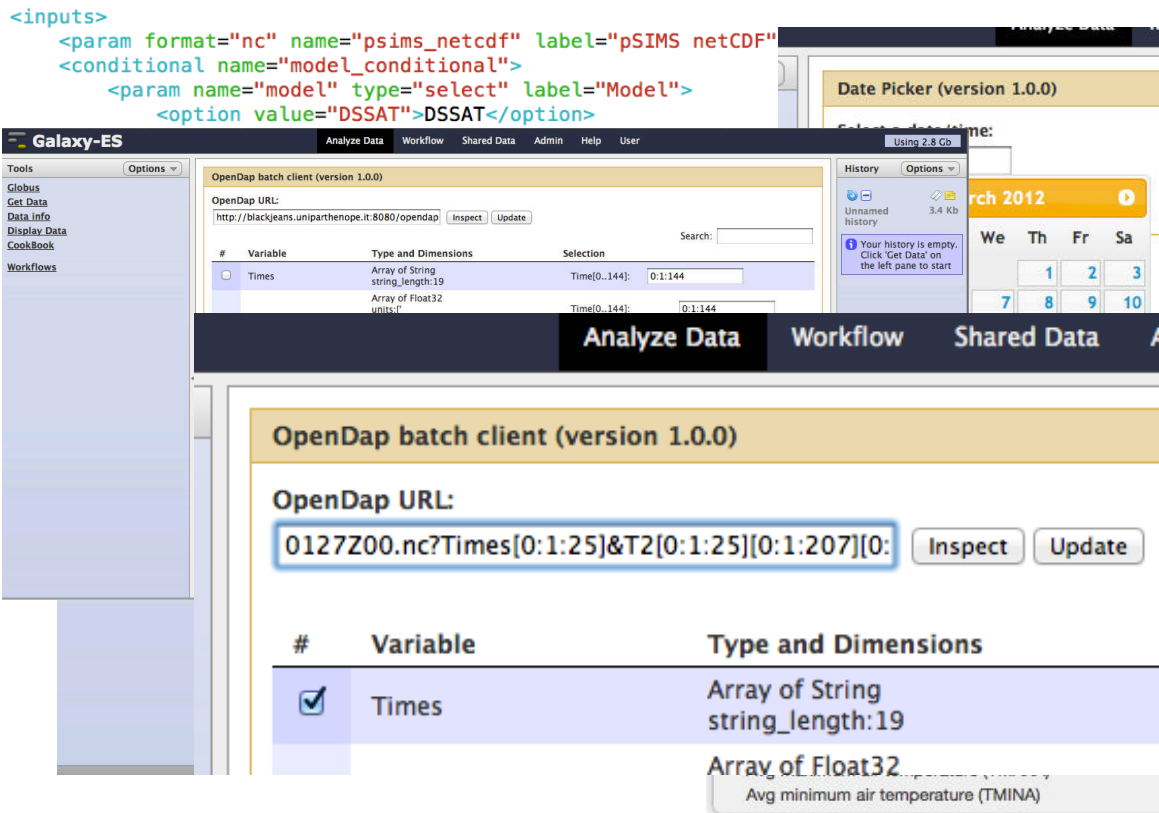
...

```
</variable>
```

```
<tool id="gcm2gcm latlon" name="GCM to GCM with latlon" version="0.1">
  <description>Convert a GCM dataset to a GCMlatlon ready for WMS ...</description>
  <command>ncpdq -a lat,lon $Input $Output</command>
  <inputs>
    <param name="Input" type="data" format="gcm.nc" label="..." />
  </inputs>
  <outputs>
    <data format="gcm.latlon.nc" name="Output" label="..." />
  </outputs>
</tool>
```

- **Tool parameters:**
Define the user interface elements for a tool
- Regular tool parameters wrap text fields, radio buttons and drop down lists.
- Custom tool parameters for Globus Online, OpenDap, date peaking and feature selection of maps.

```
<inputs>
<param format="nc" name="psims_netcdf" label="pSIMS netCDF"
<conditional name="model_conditional">
  <param name="model" type="select" label="Model">
    <option value="DSSAT">DSSAT</option>
```



Galaxy-ES

Analyze Data Workflow Shared Data Admin Help User

Tools Options

Globus
Get Data
Data Info
Display Data
CookBook
Workflows

OpenDap batch client (version 1.0.0)

OpenDap URL:
http://blackjeans.uniparthenope.it:8080/opendap/ Inspect Update

Search:

#	Variable	Type and Dimensions	Selection
<input type="checkbox"/>	Times	Array of String string_length:19	Time(0..144): 0:1:144
		Array of Float32 units:1	Time(0..144): 0:1:144

History Options

Using 2.8 Gb

3.4 Kb

Unnamed history

Your history is empty. Click 'Get Data' on the left pane to start

March 2012

We Th Fr Sa

1 2 3

7 8 9 10

Analyze Data Workflow Shared Data

OpenDap batch client (version 1.0.0)

OpenDap URL:
0127Z00.nc?Times[0:1:25]&T2[0:1:25][0:1:207][0: Inspect Update

#	Variable	Type and Dimensions
<input checked="" type="checkbox"/>	Times	Array of String string_length:19
		Array of Float32

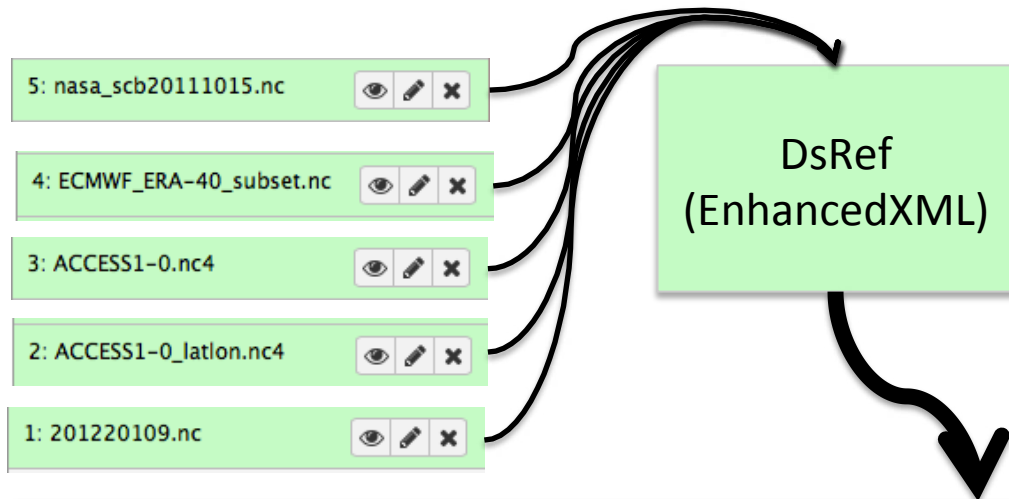
Avg minimum air temperature (TMINA)

Step FOUR: aggregated datatypes (RAFT*)

- **Dataset References:**

XML based datatype grouping references to different datasets in the same history.

- The regular Galaxy works on single file datasets or composite file datasets.
- Acts as a 'struct' or an 'array' or a mix of both.
- Supports schemas and translators.

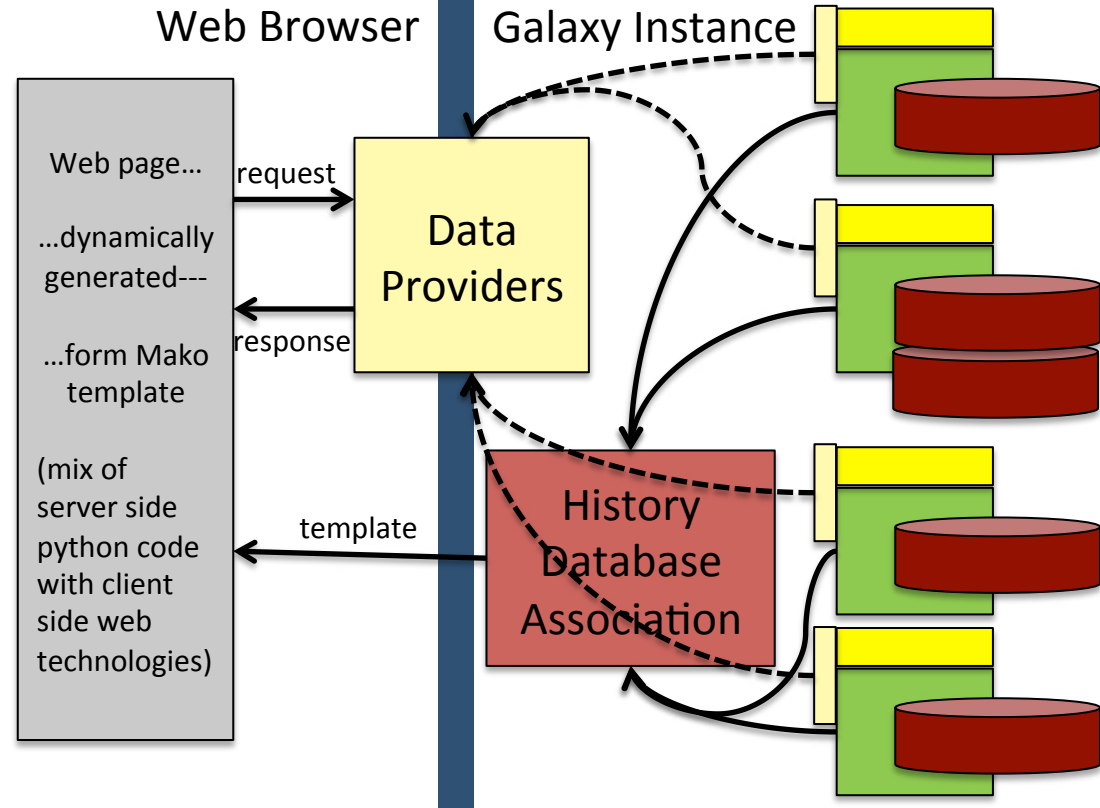
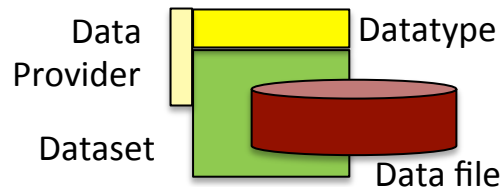


Used when:

- A tool consumes and/or produces a variable number of datasets
- The tool is implemented using a Swift script working in parallel

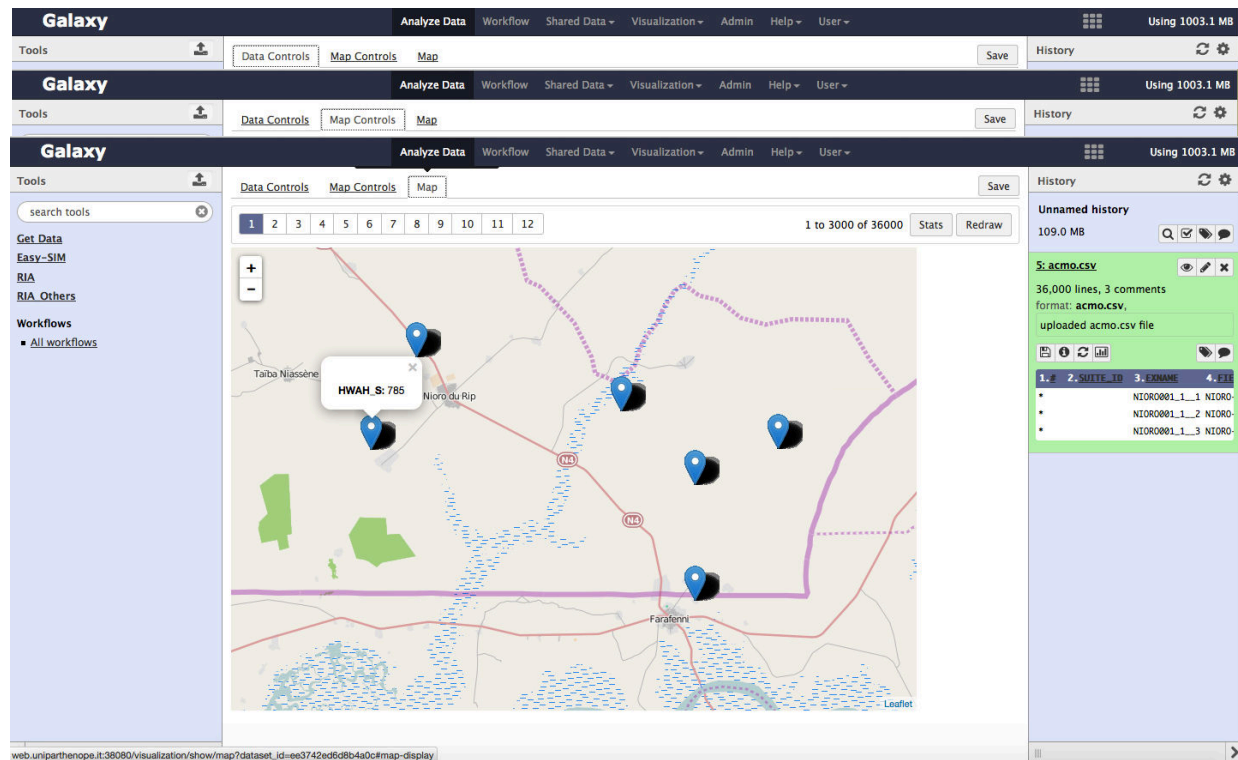
Step FIVE: data providers

- **Data providers:** software components interfacing the datasets with the web browser.
- They provide data as array of JSON objects
- Key/Values, Columnar, custom
- Implemented in Datatype classes



Step SIX: [GeoJson vector maps] map visualizers

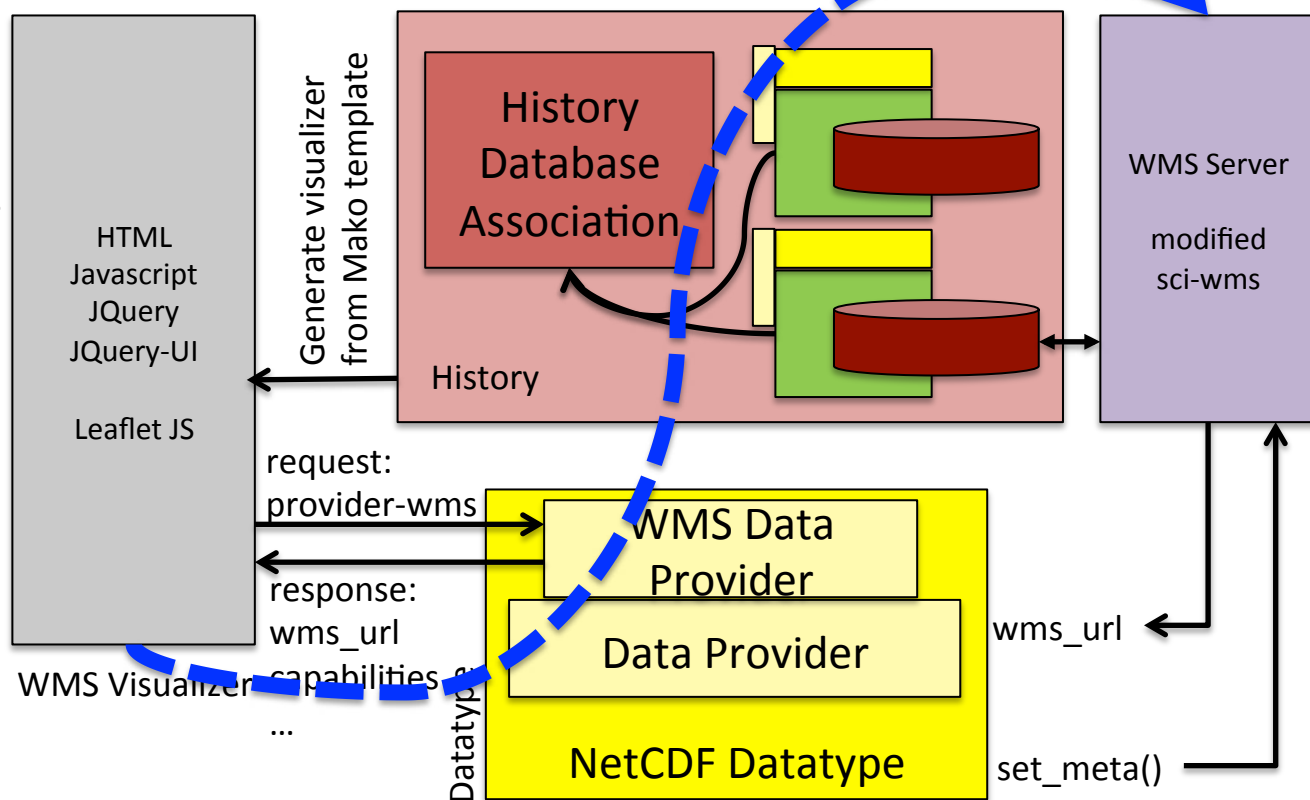
- **Visualizers:**
client-side software components for interactive data visualization
- Quasi-GIS!
- **Map:**
Visualizes vector data produced as GeoJson objects by a data provider
- **Wms** (World Map Server):
Visualizes raster data from NetCDF datatypes.



ACMO file visualization

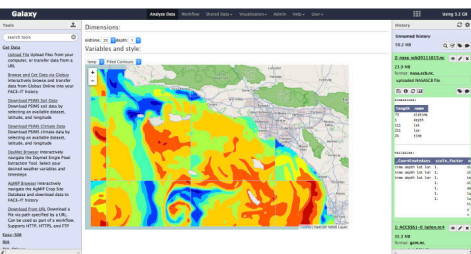
[NetCDF & World Map Server] map visualizers

- **Wms:**
World Map Server visualizes raster data from NetCDF datatypes.
- It leverages on an external software.
- Still experimental!
- Steps:
 - Dataset registration
 - Data provider interaction
 - GUI setup
 - Map consuming

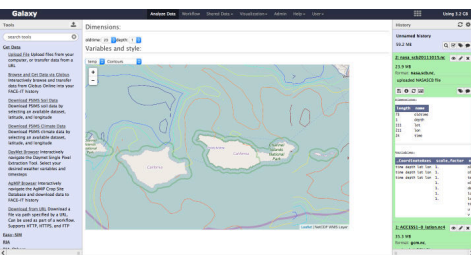


Step SIX: [NetCDF & World Map Server] map visualizers

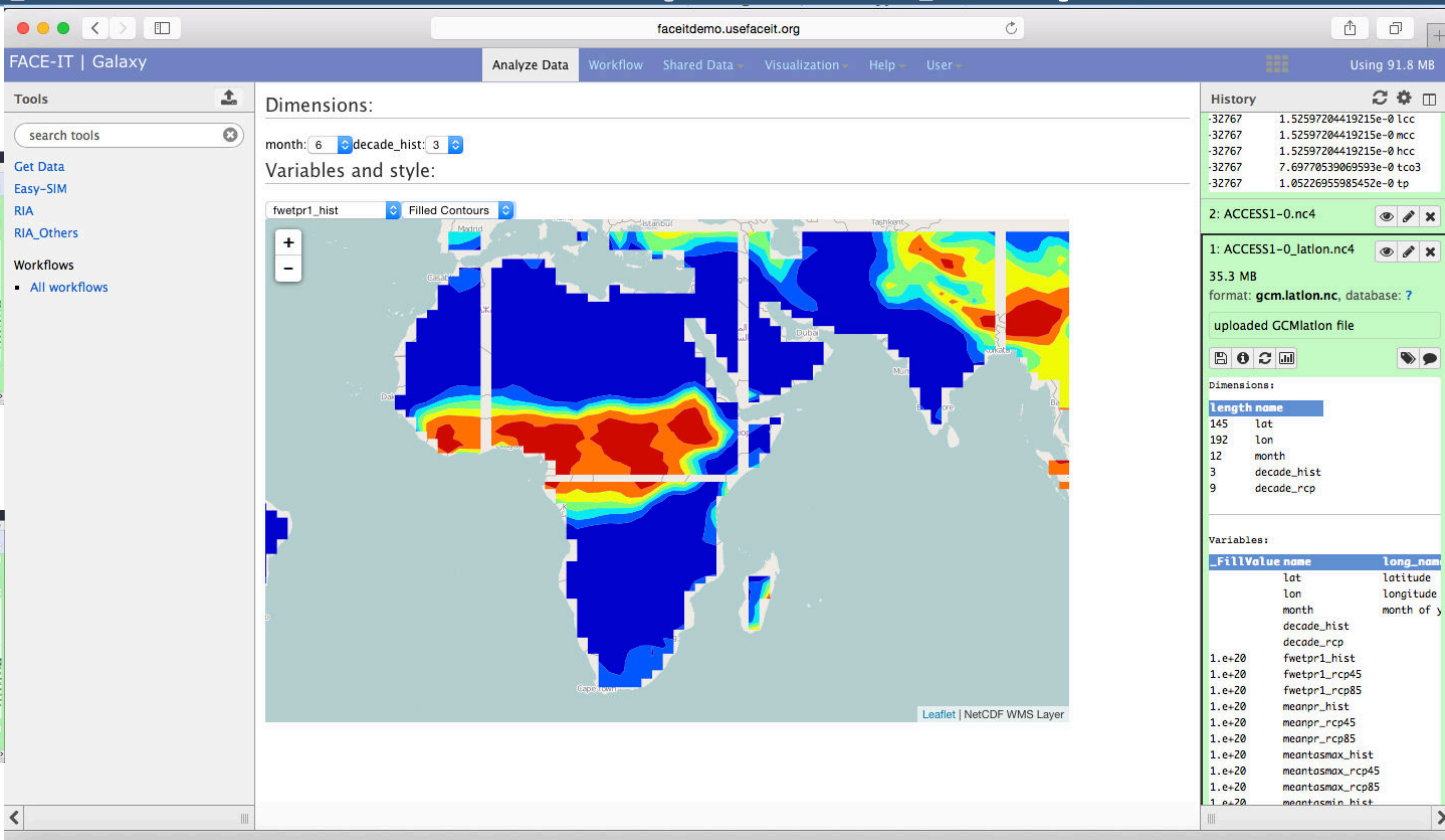
Examples:



Sea Surface Temperature



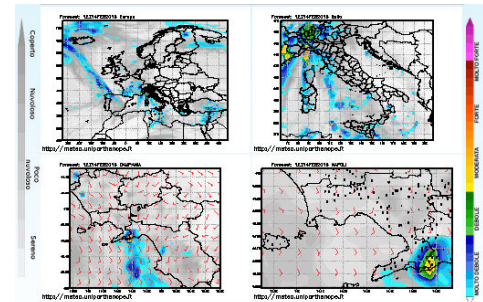
Conturing



- Face-IT Galaxy is a creative playground for the next generation of earth scientists

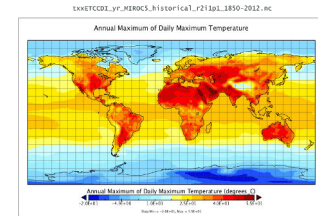
<http://www.learnfaceit.org>

- **Propose** your application, write your code and **share it!**
- Spin-off projects: extreme weather simulations in the Bay of Napoli, IT (UniParthenope)



A possible CI collaboration?

- Instrumented **Smart Cities** are a huge source of big data
- Array of Things as a Face-IT Galaxy data source?
- Why not use NetCDF as a search criteria after a crawler has explored the internet hunting for earth system data?



GCE: The 9th Gateway Computing Environments Workshop@SC14



FACE-IT: A Science Gateway for Food Security Research

Raffaele Montella^{1,2}, Alison Brizius², Joshua Elliott², David Kelly², Ravi Madduri^{1,3}, Ketan Maheshwari¹, Cheryl Porter⁴, Peter Viller², Michael Wilde², Wei Xiong⁴, Meng Zhang⁴ and Ian Foster^{2,3,5}

¹Department of Science and Technologies, University of Naples Parthenope, Naples, ITALY;

²Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, Illinois, USA;

³Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA;

⁴University of Florida, Department of Agricultural and Biological Engineering, Gainesville, Florida, USA;

⁵Department of Computer Science, University of Chicago, Chicago, Illinois, USA;



The FACE-IT Approach

Understanding the potential impacts of climate change and the likely effectiveness of adaptation strategies is of crucial importance to the sustainability of both agriculture and natural ecosystems.

Framework to Advance Climate, Economic and Impact Investigations with Information Technology (FACE-IT)

Motivations

- Multiplicity of data formats
- Inadequate computational tools
- Difficulty in sharing data and programs
- Lack of incentives for pro-social behavior
- Large data

FACE-IT Features

- Data store
- Rich program collections for format conversion
- Convenient data and code ingest mechanisms
- Rich social elements to incentivize contributions

Technologies and Tools

FACE-IT builds on the Globus Galaxy-ES platform. The Globus Galaxy-ES platform leverages Galaxy

- Simple, uniform, and extensible interface for selecting and executing components and workflows
- Globus transfer for data movement
- Globus Nexus for identity, group and profile management
- Swift for parallel execution of workflow components in large ensemble simulations
- Custom elements for elastic, scalable cloud execution

Improvements

- Appropriate Galaxy datatypes
- Earth system data providers direct integration
- RAFT for data aggregation

- NetCDF Schema
- Time series displays and visualizers
- Interactive
- map data rendering

Architecture and Implementation

Thanks! Questions?

Application within AGMIP

AgMIP climate team researchers prepare a set of historical weather data and future climate model projections, plus software tools for generating time-series weather scenarios from these inputs to be used to drive crop models.



Workflows include multiple models and produce a variety of browser visualizations and publication quality images that can be downloaded directly.

Hosted on Amazon cloud for on demand access, scalability

Builds on widely used Galaxy Globus and Swift systems

faceit-portal.org