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

Raffaele Montella\textsuperscript{1,2}, Alison Brizius\textsuperscript{2}, Joshua Elliott\textsuperscript{2}, David Kelly\textsuperscript{2}, Ravi Madduri\textsuperscript{2,3}, Ketan Maheshwari\textsuperscript{3}, Cheryl Porter\textsuperscript{4}, Peter Vilter\textsuperscript{2}, Michael Wilde\textsuperscript{2}, Wei Xiong\textsuperscript{4}, Meng Zhang\textsuperscript{4} and Ian Foster\textsuperscript{2,3,5}

\textsuperscript{1}Department of Science and Technologies, University of Naples Parthenope, Naples, ITALY;
\textsuperscript{2}Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, Illinois, USA;
\textsuperscript{3}Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA;
\textsuperscript{4}University of Florida, Department of Agricultural and Biological Engineering, Gainesville, Florida, USA;
\textsuperscript{5}Department of Computer Science, University of Chicago, Chicago, Illinois, USA;

faceit-portal.org
learnfaceit.org
usefaceit.org
FACE-IT: A Framework to Advance Climate, Economic, and Impact Investigations with Information Technology

Facing real problems with Information Technology

Data + Workflows = Results

What’s in a name...
No buzzword
Real things!
An open playground for the next generation of earth system scientists

The user profile...
Scientists
Experts of their fields
Limited programming skills
Complex experiments

Effective and efficient solutions to real problems
Experts in design and abstraction

Built on widely used Galaxy, Globus, and Swift systems

Development experts in wizardry...
faceit-portal.org
The Face-IT Galaxy timeline

Galaxy-ES is a pluggable toolshed of Globus Galaxy Genomics datatypes, more tools, more working demos. The Face-IT team is finally joined.

- Many data sources
- Many applications
- Globus integration
- People around the world use it!
- Aggregated datasets (RAFT)

2011 - 2015

Face-IT
Galaxy-ES

Face-IT at GCE14:
- New netcdf datatype with schema
- New xml and json datatypes
- WMS map visualizer

Third Face-IT developer conference:
- NTCC: No need to touch core code
- Datatypes as "proprietary"
- New visualizers
- New netcdf datatype with schema
- New xml and json datatypes
- WMS map visualizer

First Face-IT developer conference:
- Research prototype
- Galaxy-ES changes
- More datatypes, more tools, more working demos.

Second Face-IT developer conference:
- Globus Online integration
- 3rd parts remote data browsing
- Advanced visualization

Fourth Face-IT developer conference:
- AgMIP meeting.
- Many data sources
- Many applications
- Globus integration
- People around the world use it!

Changes in status:
- Research prototype to project in progress.
- The Face-IT team is finally joined.
- First Face-IT developer conference.
FACE-IT: A Framework to Advance Climate, Economic, and Impact Investigations with Information Technology
The Hitchhiker’s [Workflow] Guide to the Galaxy

FACE-IT: A Framework to Advance Climate, Economic, and Impact Investigations with Information Technology
FACE-IT: A Framework to Advance Climate, Economic, and Impact Investigations with Information Technology

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

From genes to Earth
**Step ONE:**

**earth system datatypes**

<table>
<thead>
<tr>
<th>Data</th>
<th>NetCDF</th>
<th>GCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>file_ext</td>
<td>file_ext=&quot;nc&quot;</td>
<td>file_ext=&quot;gcm.nc&quot;</td>
</tr>
<tr>
<td>mime-type</td>
<td>Metadata: NCML</td>
<td>Metadata: NCML</td>
</tr>
<tr>
<td>Metadata</td>
<td>Metadata: WMS</td>
<td>NCML_Metadata: WMS</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **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

FACE-IT: A Framework to Advance Climate, Economic, and Impact Investigations with Information Technology
• **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
**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
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2">
  <dimension name="lat" length=",*" />
  <dimension name="lon" length=",*" />
  <variable name="meantasmmax_rcp45" shape="decade_rcp month lon lat" type="float">
    <attribute name="_FillValue" type="float" value="1.e+20" />
  </variable>
  <variable name="meantasmmax_rcp85" shape="decade_rcp month lon lat" type="float">
    <attribute name="_FillValue" type="float" value="1.e+20" />
  </variable>
</netcdf>
```
Step TWO: new tools

- **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
Step TWO:
[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

```xml
<tool id="gcm2gcmlatlon" 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>
```
Step THREE: tool parameters

- **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.
**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
**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

Web page...
...dynamically generated---
...form Mako template
(mix of server side python code with client side web technologies)
**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.
Step SIX: 

[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
FACE-IT: A Framework to Advance Climate, Economic, and Impact Investigations with Information Technology

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)
Conclusions and [tomorrow] future works

- 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?
FACE-IT: A Science Gateway for Food Security Research

Raffaello Montella, Alis on Brizius, Joshua Elliott, David Kelly, Ravi Madduri, Ketan Maheshwari, Cheryll Porter, Peter Viteri, Michael Witte, Wei Xiong, Meng Zhang, and Ian Foster

1Department of Science and Technologies, University of Naples, Naples, ITALY;
2Computer & Information Science Division, Argonne National Laboratory, Argonne, Illinois, USA;
3University of Florida, Department of Agricultural and Biological Engineering, Gainesville, Florida, USA;
4Department of Computer Science, University of Chicago, Chicago, Illinois, USA;
5Department of Agriculture and Engineering, University of Chicago, Chicago, Illinois, USA;
6Department of Mathematical and Computer Science, University of Naples, Naples, Italy.

GCE: The 9th Gateway Computing Environments Workshop@SC14

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

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 ingestion 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 for 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 data types
- Earth system data providers
- Direct integration
- RAMP for data aggregation
- NetCDF Schema
- Times series displayers and visualizers
- Interactive maps
- Data rendering

Application within AGMIP

AGMIP climate team researchers prepare a set of historical weather data and future climate model projections, and project them to a time series weather scenarios from these inputs to the used to drive crop models.

Results

A comparative advantage study of the impacts of climate change on crop yields in the Great Plains.

Thanks! Questions?