A Collaborative Platform for Integrating AgroInformatics Data Using Globus

Andrew Gustafson
Minnesota Supercomputing Institute
University of Minnesota

Team: Philip Pardey, Jim Wilgenbusch, Kevin Silverstein, Getiria Onsongo, Michael Milligan, Tom Prather, Ying Zhang

AgroInformatics:

Using agricultural bioinformatics data sets to understand factors affecting crop performance, in order to improve agricultural outcomes.

Some Questions of Interest:

- Which crop varieties are most resistant to pests?
- Which crop varieties do best in particular climates?
- Are there genetic commonalities between crops with desirable resistive or productive features?
- What factors are most important in predicting farm productivity?

The Goal:

To create a platform and toolset to allow researchers to upload and analyze agroinformatics data, and share data sets with each other in a controlled way.

The International AgroInformatics Alliance (IAA)

An alliance of institutions seeking to build the platform, and share accumulated datasets.

The platform is hosted at the
Minnesota Supercomputing Institute (MSI).

MSI Computing and Data Storage Assets

- 6 PB of global storage; 3 PB of Tier2 storage; Spectralogic T950 “Archive” + backup
- Mesabi: HPC System; 774 nodes; >18000 cores; Infiniband; Large Memory; GPUs
- Itasca: Circa 2010 system; 800 nodes; >6000 cores
- Virtualization Resources; Interactive Computing; Hadoop Cluster
- Networking 2 x 10 GBE; 100 GBE soon. Globus data transfer nodes.

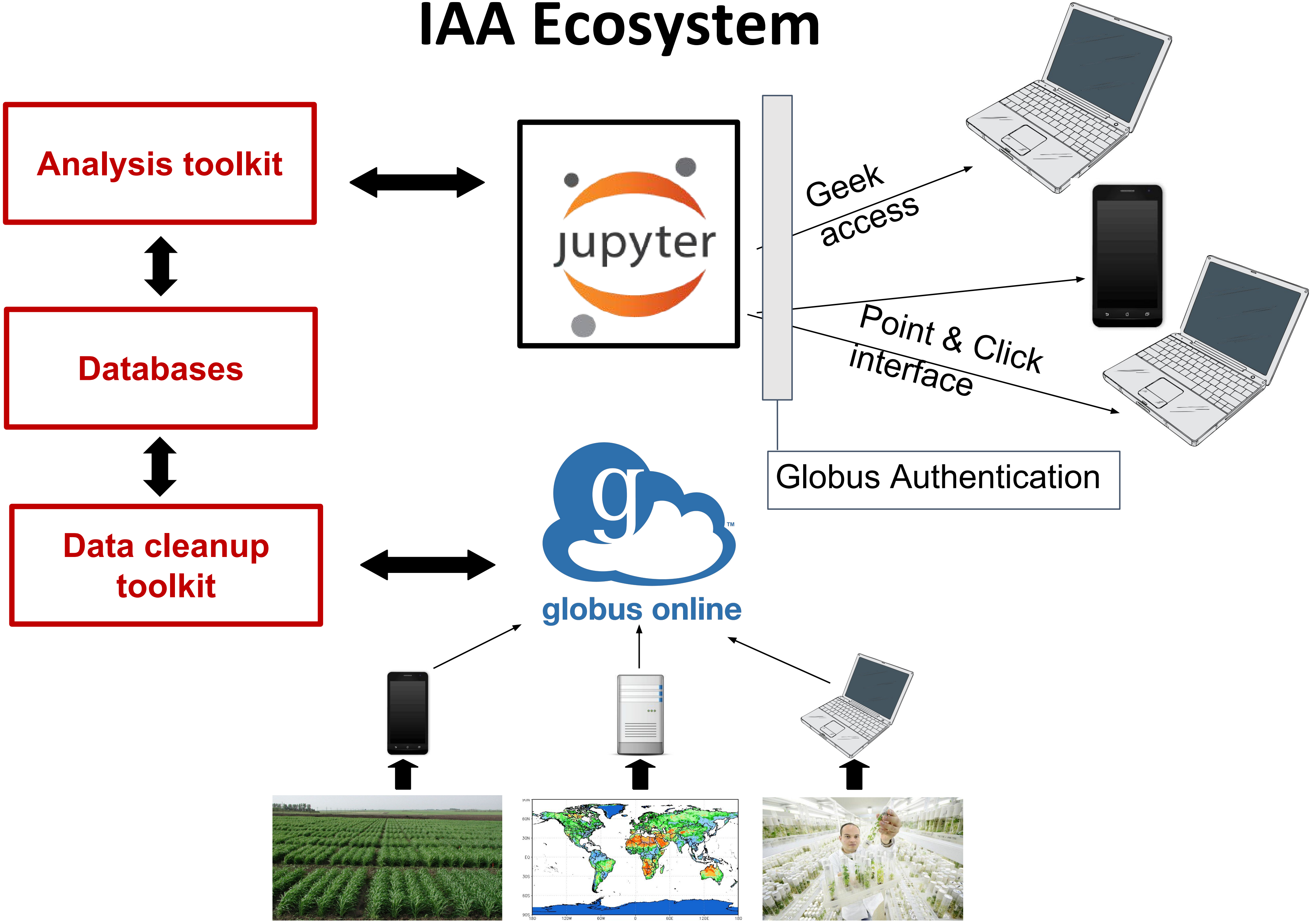
Platform Design Goals (Part I)

- Data Storage
  - Reliable database with a reputation for data integrity and correctness
  - Curating & aligning new + previously isolated data sets
- Data Security
  - Authentication: use other sources
    - Partner universities, Google, etc.
  - Authorization: Three levels of access controls
    - Single organization
    - Defined set of people/users
    - Everyone who is part of the Alliance

Platform Design Goals (Part II)

- Data Transfer
  - Encryption in flight should be available
  - Take advantage of high speed networks
  - Accommodate slow and unreliable connections
- Analysis Platform
  - Leverage existing software libraries & hardware technologies
  - Accommodate a variety of analysis styles
  - Accommodate a variety of programming languages
  - Accommodate a wide-range of technical expertise

IAA Ecosystem



Analysis toolkit

Databases

Data cleanup toolkit

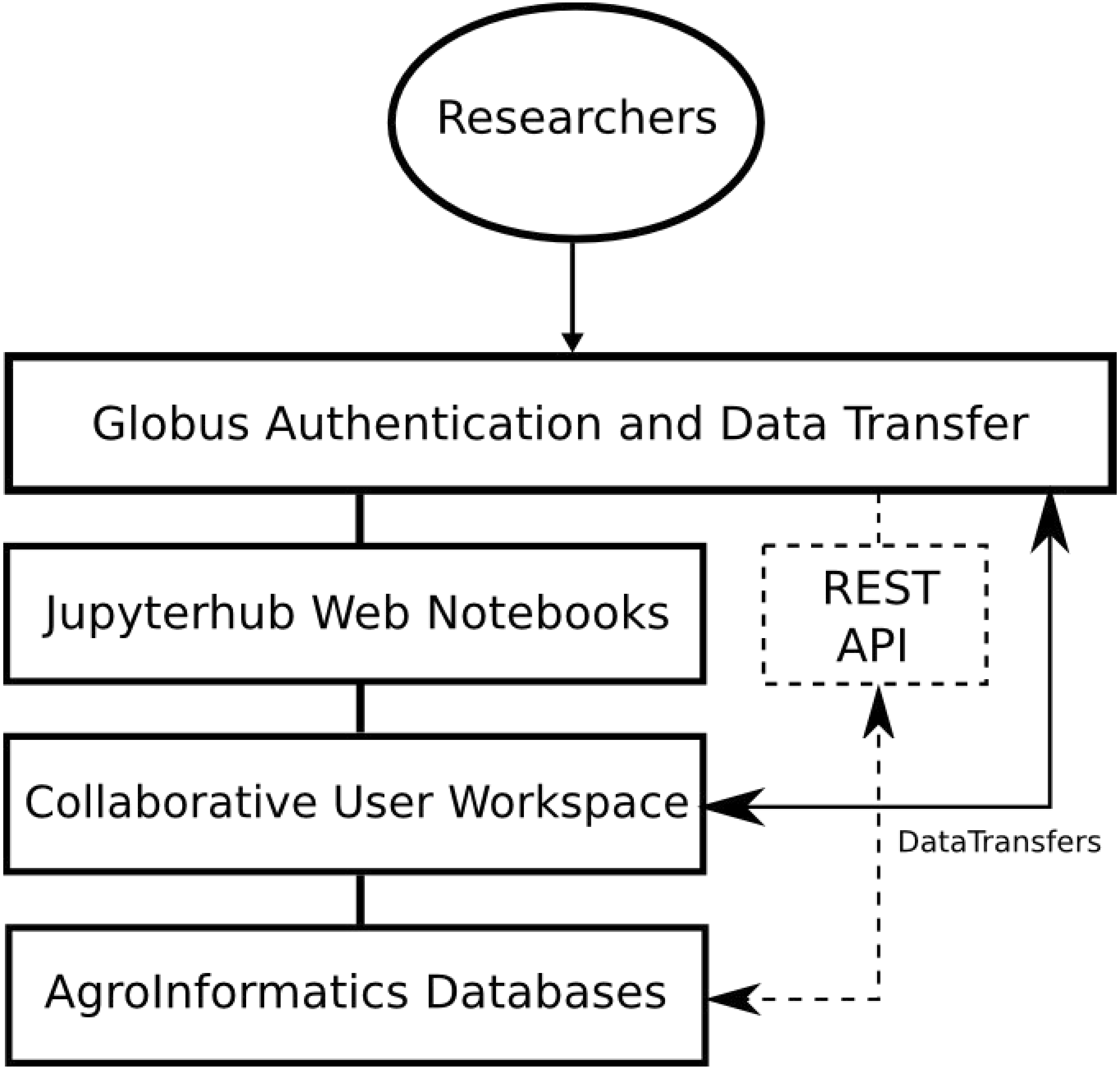
jupyter

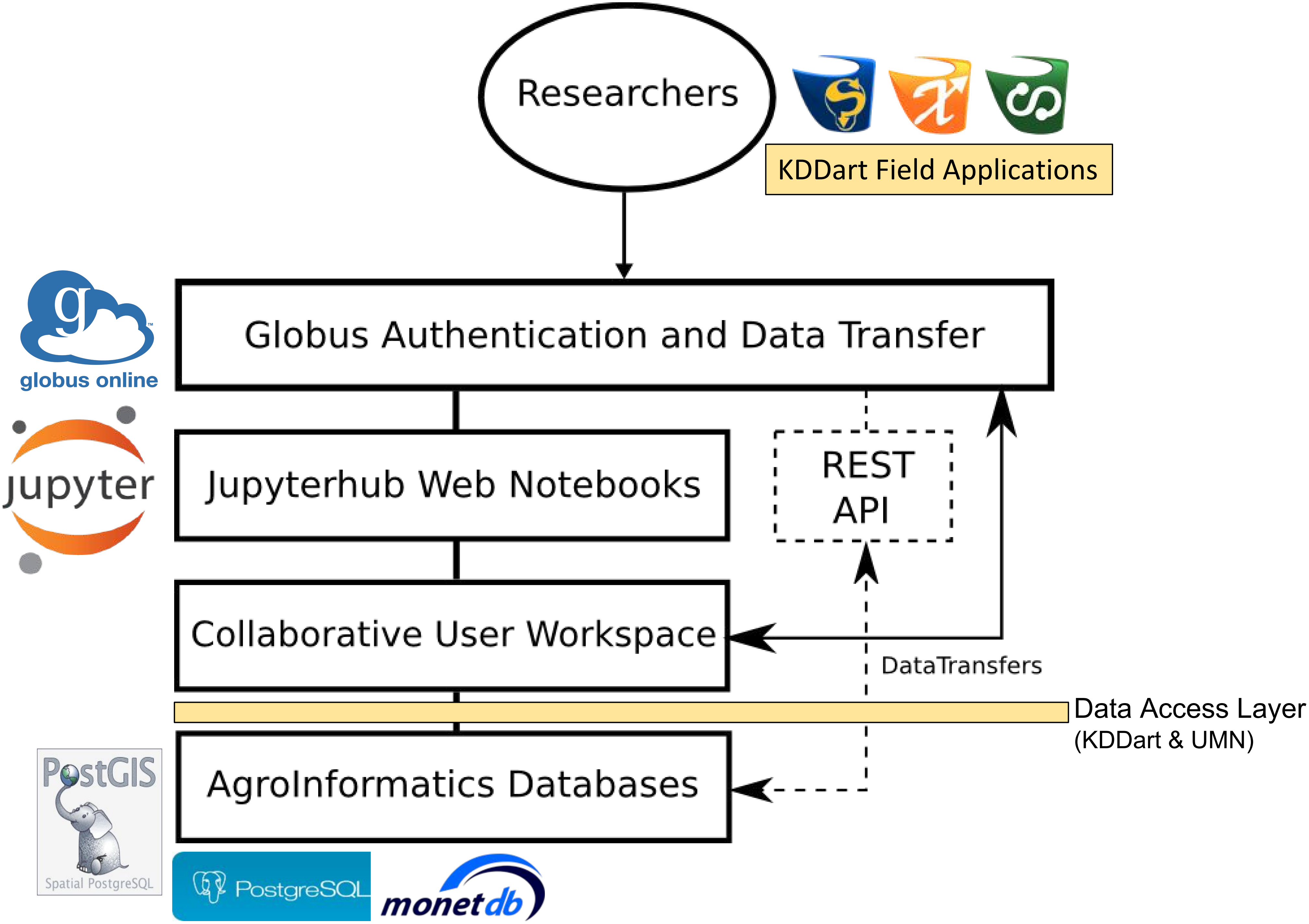
Geek access

Point & Click interface

Globus Authentication

globus online


Researchers
Globus Authentication and Data Transfer
Jupyterhub Web Notebooks
REST API
Collaborative User Workspace
DataTransfers
AgroInformatics Databases



Researchers

KDDart Field Applications

globus online

Globus Authentication and Data Transfer

jupyter

Jupyterhub Web Notebooks

REST API

Collaborative User Workspace

DataTransfers

Data Access Layer (KDDart & UMN)

AgroInformatics Databases

PostGIS Spatial PostgreSQL

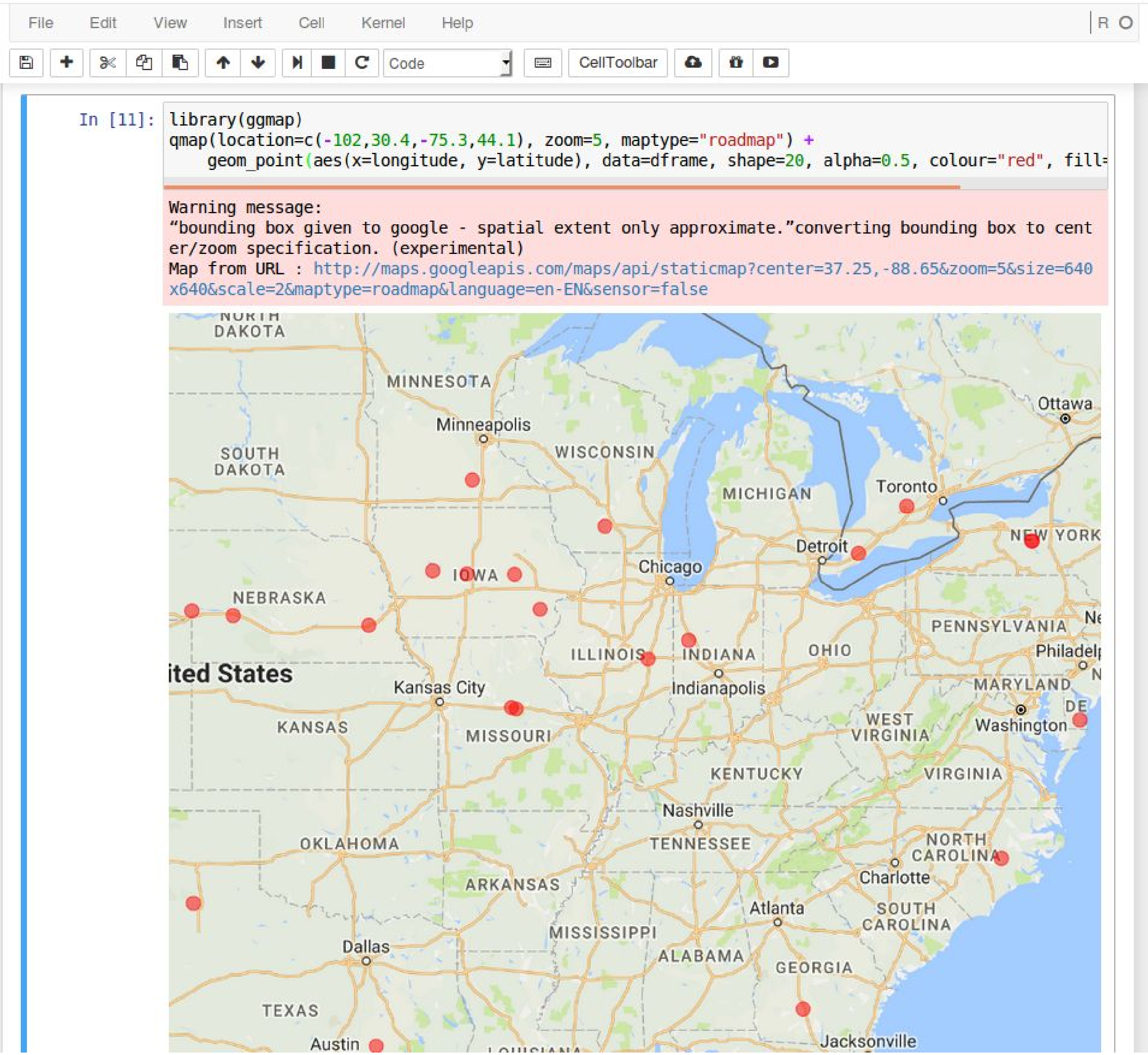
PostgreSQL

monetdb

Platform: Key Elements

- Authentication – Globus
  - How we confirm that you are who you say you are.
- Data Authorization – IAA and KDDart
  - Controls what assets you can see
- Data Transfers – Globus and KDDart
  - Globus acts as a robust service layer for moving data over high-speed networks
  - KDDart allows for data movement using mobile and others
- Data Storage – PostgresSQL, PostGIS, MonetDB
  - PostGIS is a secure spatial dataBase that extends PostgreSQL
- Analysis Environment – Jupyter, Web, and KDDart
  - Support for constrained (point-click) and unconstrained (Python and R) analysis environments

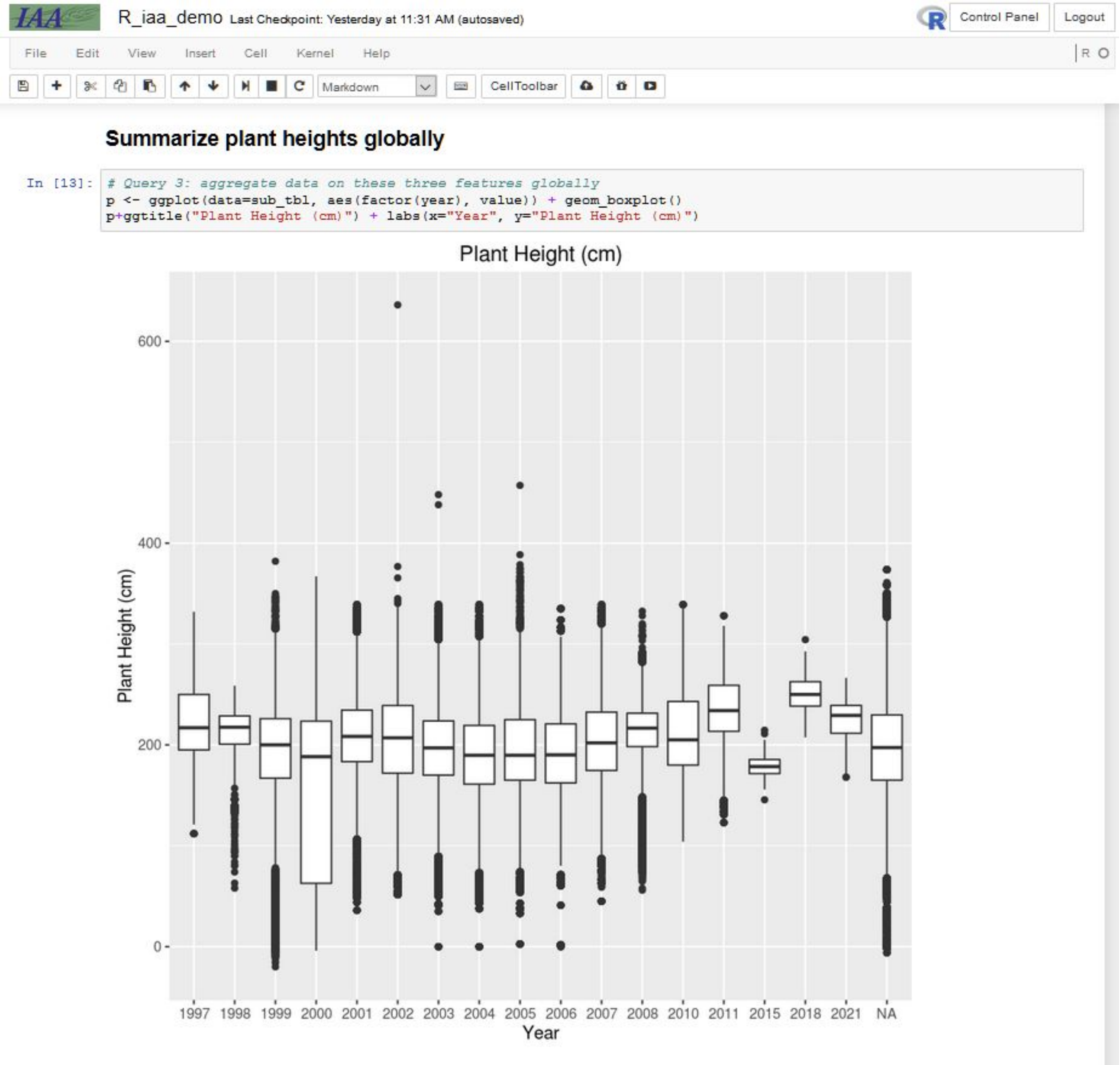
Snapshot of Live Geek interface (1)



```
In [11]: library(ggmap)
         qmap(location=c(-102,30.4,-75.3,44.1), zoom=5, maptype="roadmap") +
             geom_point(aes(x=longitude, y=latitude), data=dframe, shape=20, alpha=0.5, colour="red", fill=
```

```
Warning message:
“bounding box given to google - spatial extent only approximate.”converting bounding box to cent
er/zoom specification. (experimental)
Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=37.25,-88.65&zoom=5&size=640
x640&scale=2&maptype=roadmap&language=en-EN&sensor=false
```

Snapshot of Live Geek interface (2)

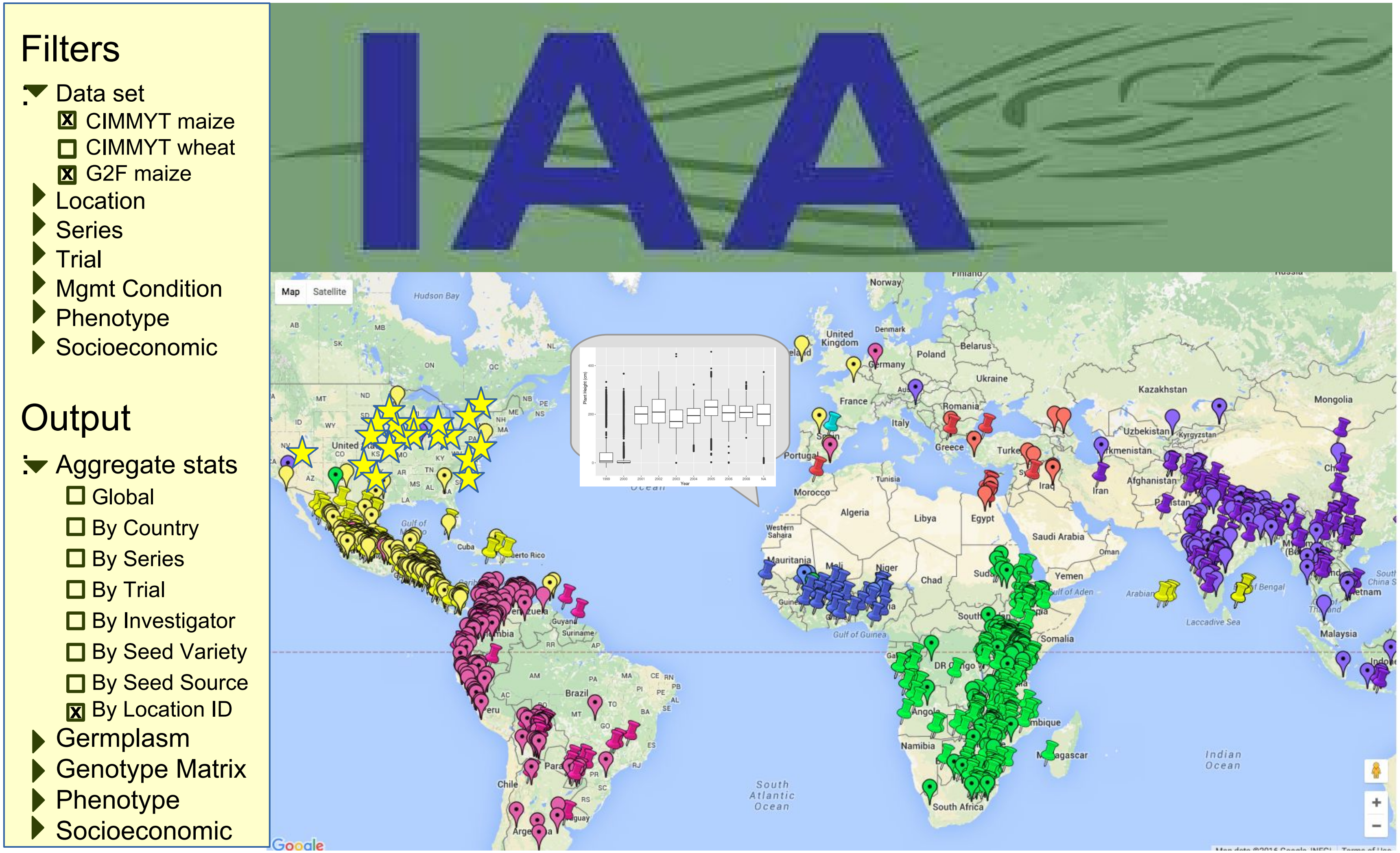


Summarize plant heights globally

In [13]:

```
# Query 3: aggregate data on these three features globally
p <- ggplot(data=sub_tbl, aes(factor(year), value)) + geom_boxplot()
p+ggtitle("Plant Height (cm)") + labs(x="Year", y="Plant Height (cm)")
```

Prototypes of point-and-click interface



Filters

- ▼ Data set
  - ☒ CIMMYT maize
  - ☐ CIMMYT wheat
  - ☒ G2F maize
- ▶ Location
- ▶ Series
- ▶ Trial
- ▶ Mgmt Condition
- ▶ Phenotype
- ▶ Socioeconomic

Output

- ▼ Aggregate stats
  - ☐ Global
  - ☐ By Country
  - ☐ By Series
  - ☐ By Trial
  - ☐ By Investigator
  - ☐ By Seed Variety
  - ☐ By Seed Source
  - ☒ By Location ID
- ▶ Germplasm
- ▶ Genotype Matrix
- ▶ Phenotype
- ▶ Socioeconomic

Mousing over a location displays selected aggregate stats for that location.

Current Platform State

A first version is built which contains:

- Interactive analysis spaces using Jupyterhub notebooks (supporting Python and R environments).
- Databases hosting a variety of data types (genomic, spatial, etc.).
- Data cleanup and analysis tools.
- Globus integration for data transfer and authentication.

A second version is being constructed which will include a “point-and-click” user friendly interface, a REST API for convenient remote queries, and more...