# Geoanalytics

TR-11-03
Jeff R. Heard
July 25, 2011

# Geoanalytics

*A cyberinfrastructure for geography applications to science*

## Background

The pressures of producing science with global relevance and global impact have made understanding and using geographic essential to a large portion of scientific research. Geographic information systems live at the heart of projects in public health, environmental science, policy and government, situational awareness, and others. Geographic data need has also become a Big Data need. Datasets essential to these projects are often terabytes in size, or they are rapidly evolving streams of complex data.

The tools available to professionals looking to *do* things with geographic data have not grown to meet the Big Data problem. Traditional GIS allows researchers to build custom databases with analytics. Google Maps allows them to publish data to the web. Various open source tools exist for specialized and general GIS purposes. As of yet, the world is without a compelling infrastructure for integrating these. As a result, geographic solutions to scientific and social problems are often cobbled together, and the results are isolated *silos* that cannot be easily integrated or adapted to new and different data.

Traditional GIS solutions like ArcGIS and GRASS allow a user to import a number of maps and work with them as a project, doing complex analysis, but the results of this are offline, or at very least relatively static. There exist "onlining" modules for these, but they are built on a pre-web paradigm whose thinking pervades the online experience. Modern users expect integration, mashups, and web-based application platforms that include data "at the bleeding edge of now."

What we will call "second generation" solutions, like Google Maps and Google App Engine allow a user to quickly create a map without prior training, requiring only a text editor, a web browser, and some patience. These solutions, in their simplicity, abstract away functionality that is needed for serious scientific analysis. These must be completed with other tools, imported into data formats tailored towards visual presentation, and do not preserve the data for analysis. This adds complexity to the scientific process as well as discourages the sharing of source data.

## Intent

We present our vision for a third-generation solution, RENCI's Geoanalytics Cyberinfrastructure, to the working with geographic data. This system will be developed according to real-world needs arising in domain sciences, and will be directed toward the following goals:

- Scale horizontally to Big Data, its update frequency, access patterns, and management.
- Integrate sensible data management solutions to scale.
- Vet and federate FOSS tools to lower the barrier-of-entry to using big geographic data.
- Provide pathways to accomplish common tasks.
- Allow a user to rapidly develop and deploy prototypes and finished solutions.

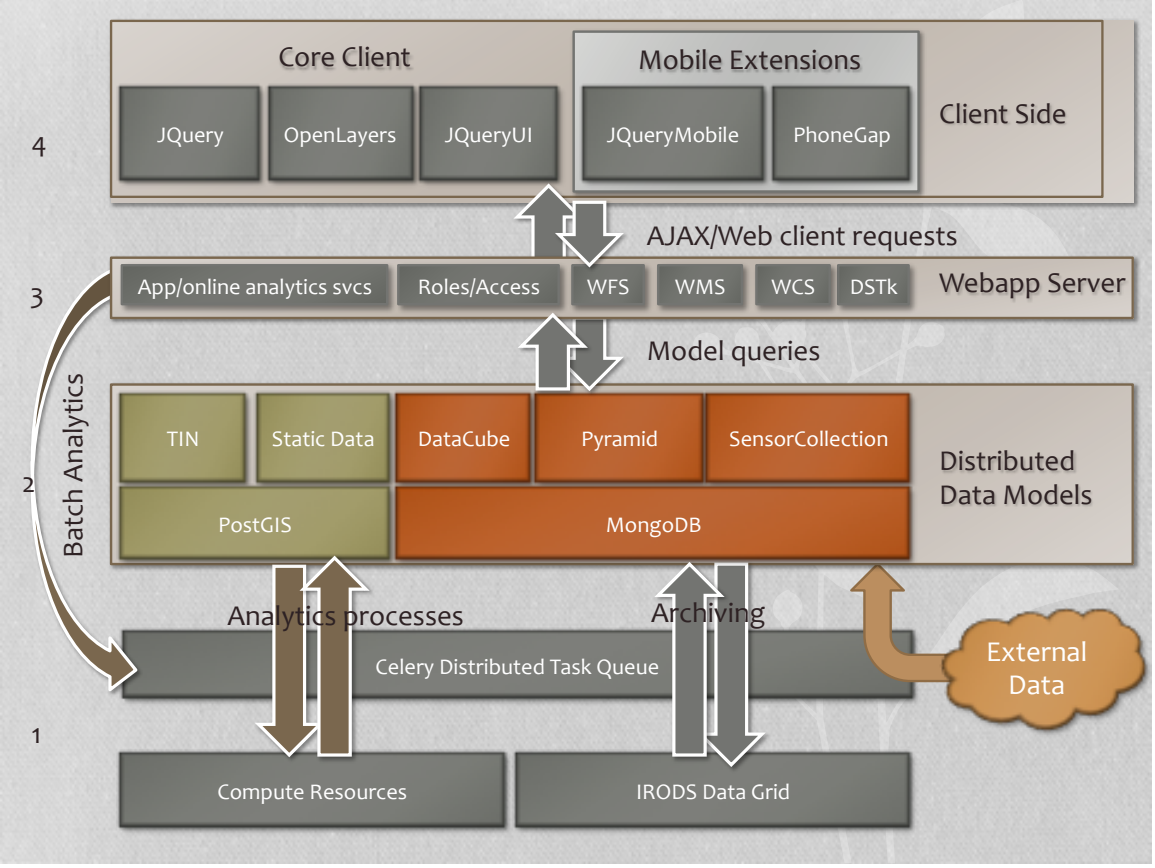


**Figure 1 Architecture of Geoanalytics**

The result of building a system to meet these goals is Geoanalytics, a software ecosystem that provides modular services for working with geographic data. Broadly, Geoanalytics' architecture breaks down into four layers, shown in the Figure 1:

1. A data management and analytics layer incorporating iRODS, open source GIS software, our supercomputing resources, and a distributed task queue.
2. A set of distributed geographic data models that encompass common data patterns. These communicate with the grid resources through the task queue.
3. A rapid web-application development platform, based on Django and open source GIS tools that includes implementations of open web-service standards WFS, WMS, and WCS layered on top of the models.
4. A set of recommended client-side technologies that form a core client layer for rapidly developing browser or mobile web-based applications.

## Data management

"Data management" has come to mean systems that provide services for archiving, federating, and automatically processing data. The Integrated Rule-Oriented Data System (iRODS) has been developed for over 10 years and provides a number of advantages over other data management systems.

These include he ability to encode data management policy in rules that execute transparently as the IRODS virtual file system is used. This achieves two important goals of: collocating data processing with data no matter its locality and providing insurance that policies are executed on all data in the system without human intervention. iRODS also has the ability to federate data stores across multiple domains, creating a Data Grid that can be expanded as needed.

In Geoanalytics, iRODS is used to:

- Manage a user's, system's, and application's access rights to data.
- Take data semi-offline and manage its end-of-life for datasets as they age.

RENCI's iRODS expertise is broad. We have multiple data grids online and have a group dedicated to research in data management on the iRODS platform.

To move data to and from the data grid into and out of indexed, online random-access data models, we employ a distributed task-queue that uses rules to determine how and when to retrieve data. This task queue can be used to automatically refresh data sources from remote sites, such as government data repositories, or can be used to automatically sunset data that has become obsolete, archiving it or deleting it as desired.

## Open data interchange and access standards

Critical to building large, enterprise or web-scale applications is the ability to speak common data interchange languages. GIS has traditionally faltered in this: geographic formats are almost as numerous as there are geographic projects; even single government agencies have been known produce data in multiple incompatible formats. This means that the first step in many projects including data from multiple domains is a data-import step where data sources are normalized into a single common format and projection scheme.

Recently the Open Geographic Consortium has introduced a number of standards to provide for interchange of data, and from these a number of web services have become key to providing applications with data in the format they need. Geoanalytics' goal is to support a subset of these standards that casts a wide net over common data access and interchange problems.

All standard and custom data models presented in the following sections support OGC's WMS (Web Mapping Service) for output, which provides styled visualization over web services. We extend WMS additionally to handle querying the underlying datasets and to handle in particular the time and elevation variables smoothly. Additionally, for TIN, SensorCollection, and custom models, we provide OGC's WFS (Web Feature Service) for accessing data its associated and geometry. Finally, for DataCube and Pyramid, we provide WCS functionality, which provides raster datasets containing underlying data values as opposed to the same data formatted for display.

## Distributed geographic data models

There are a few common patterns in geographic datasets:

- Tiled data pyramids, common for large sets of high resolution satellite imagery.

- Spatial or spatiotemporal volumetric datasets, commonly used environmental, climatological, and geophysical datasets.
- The spatiotemporal feature collection, commonly found with sensor data.
- The feature collection, which contains static vector data.
- The coverage, which holds static raster data.

Geoanalytics provides ready-to-use and customize solution to managing and querying these kinds of datasets and does so on a distributed platform that allows datasets to scale to large size without increasing the complexity of data access. To do this, these models hybridize relational databases and non-relational data stores (commonly known as NoSQL databases).

All models given here provide access methods designed for application programmers and web-services based on OGC standards for network access of data.

## DataCube


**Figure 2; DataCube architecture and query behaviour**

Spatio-temporal volumetric datasets are common in environmental modeling and weather prediction. They are characterized by dense rasters of 2 or 3 dimensional raster data defined over a geographical area, and over a span of time. Particularly for forecast datasets, an additional dimension of "version" may be desired that marks the prediction for a single x-y-z-t raster cell as being newer or older.

The DataCube model handles this case, providing selection of data "swatches" across x, y, z, time, and version constraints. Swatches are returned as FORTRAN style arrays to the application programmer, or can be returned as NetCDF formatted data over the Web.

## PIN

Another common environmental model is that of the "Polygonal Interconnect Network" or PIN. A PIN is a densely packed, but irregular raster similar to a data cube, where raster cells are of variable size of shape. This is common for models with variable resolutions, such as ADCIRC or SLOSH.

The PIN model handles this case, with an architecture and query mechanism similar to the DataCube, but paired with the interconnect network of X,Y,Z linkages between data points.

## Pyramid

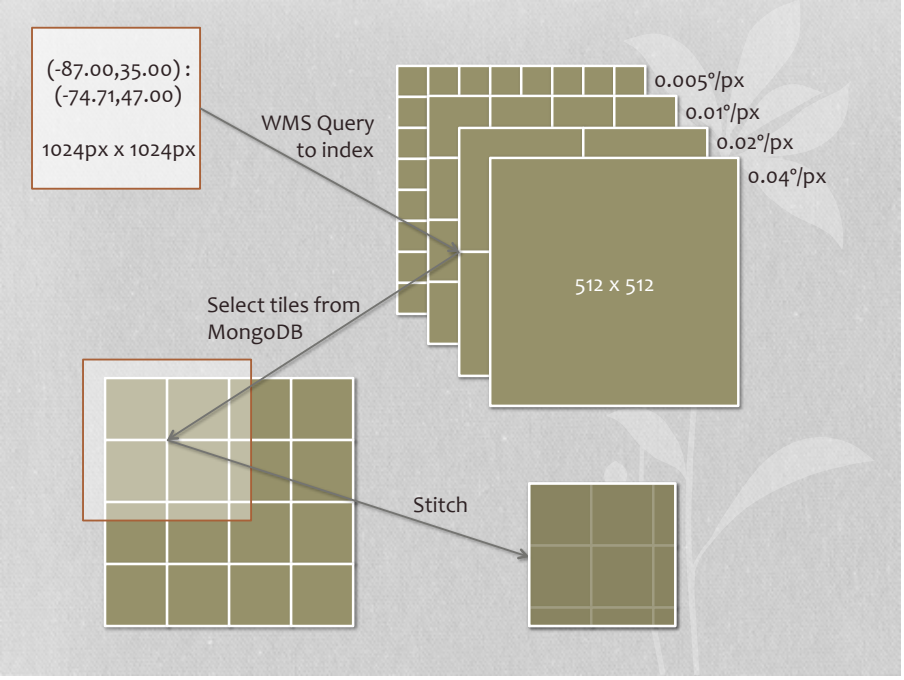


**Figure 3 Pyramid architecture and query behaviour**

When dealing with high-resolution satellite or fly-over imagery, data stores designed for online access of imagery break it up into regularly sized tiles, and then "mip-map" these tiles, creating a pyramid of tiles that can be accessed to stitch together imagery at different zoom levels quickly.

The Pyramid data model is designed to handle just this case, where a large set of imagery, such as the orthophotography for an entire state, can be processed, hosted, and indexed, and accessed in a distributed manner.

## SensorCollection

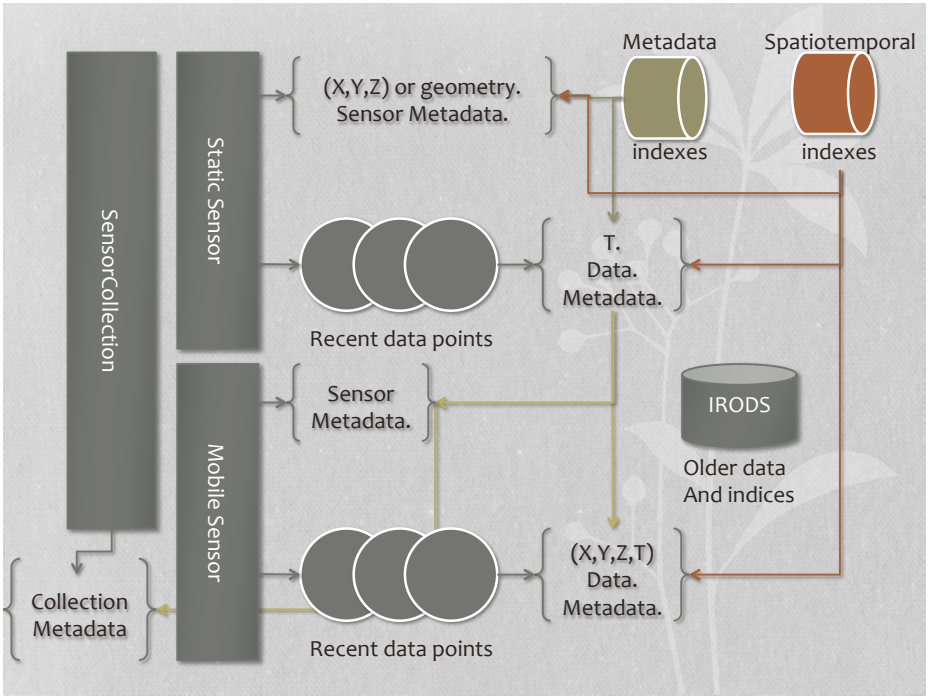


**Figure 4 SensorCollection architecture**

In 2008, there were for the first time more network-connected devices than there are people. Everything from environmental sensor networks to car alarms are being connected to the internet and their data feeds are being packaged and harnessed for newer and bigger applications. These devices can broadly be classed as "sensors."

The SensorCollection model is a flexible data model designed to aggregate data from mobile and static sensor networks with static or variable membership. That is, the SensorCollection can handle a traditional sensor network, such as an array of traffic cameras, *and* it can handle a rapidly evolving and changing mobile phone "crowdsourcing" network. SensorCollections capture not only data but metadata, and can store nested KVP metadata at the collection, sensor, or sensor-update level.

Sensors send updates with arbitrary hierarchical or regular data, and the sensor collection can stream these updates in and index based on arbitrary data parameters as well as X, Y, Z, and time.

Additionally, the SensorCollection model provides configurable strategies for reducing the "buildup" of sensor data in an online database. It provides services for archiving and sunsetting old data through iRODS as data ages or online databases become too large.

## Custom Models via GeoDjango

If none of these cases will do, then Geoanalytics provides the facility to create custom models that take advantage of all the other services Geoanalytics brings to bear. These custom models are indexed in a relational store or hybrid relational/non-relational store. As with all other models, custom data models can be used with Geoanalytics' inbuilt WMS and WFS services.

## Rapid application development

Geoanalytics' service architecture is built on top of the popular open-source Django MVC web application framework. This framework provides the scaffolding to rapidly develop and deploy web-applications that visualize and provide data stored within the geoanalytics' system.

This scaffolding includes:

- Rich stylesheets for WMS services that allow for greater control and functionality than the OGC's defined SLD standard for map styling.
- Application generation scripts that can generate a functional application on top of a data model in minutes.
- Predefined tasks for common data management operations.
- Recommended technologies to build interactive, mobile-capable interactive mapping sites that work with modern browsers.

## Analytics

"Canned" analytic routines are not a direct part of Geoanalytics' architecture. We are currently exploring incorporating a number of GRASS modules into Geoanalytics' core to cover basic analytic features, however Geoanalytics includes facilities for accessing data in serializable, programmer-friendly formats that lend themselves to processing with externally developed analytics applications. These analytical applications can be "wired into" a deployed application either with the distributed task queue, which can send data to be processed externally and wait on results, or by iRODS, which can run rules on data and retrieve and store these results.

## The Big Board

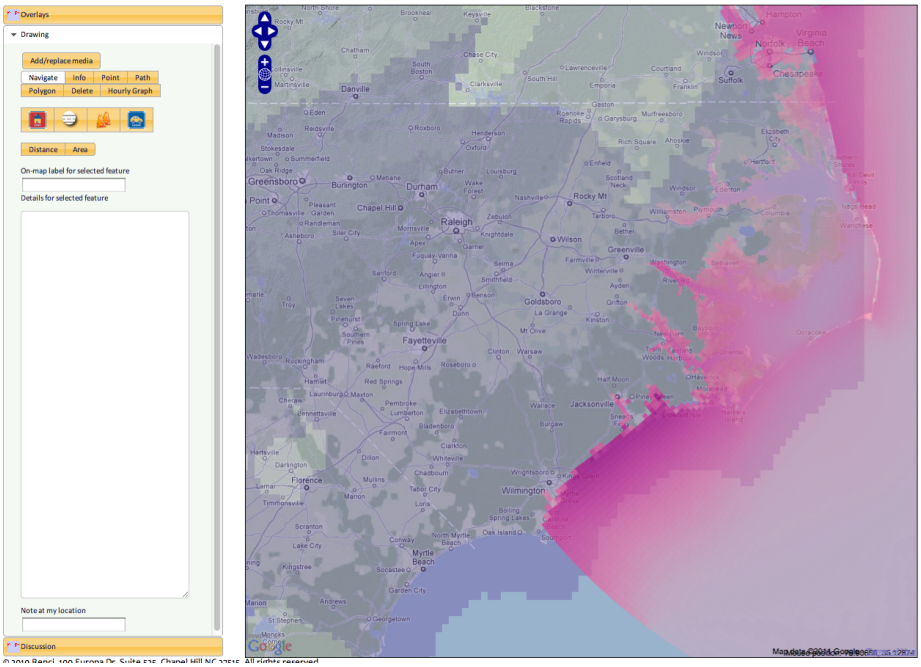


**Figure 5 The Big Board**

The Big Board is a web- and mobile-based platform for realtime map-based collaboration. It essentially provides "teleconferencing over maps." A user creates a "conference room" by selecting a base map layer and a starting location. Once created, users can join the conference room by navigating to it with their web browsers. Any number of people can be logged into the conference room at once.

The conference room page is organized into three tabs and a map. The tabs contain: overlays, drawing and analysis, and discussion. The overlay tab contains a list of overlays that can be added to the map. The overlays that can be added are determined by the role of the logged in user, and are set up ahead of time by the conference leader or administrator. Initially, when an overlay is added, it is treated as private to the user. However, these overlays each come with a "share" button that allows a user to share an overlay with all users when the data becomes relevant. Overlays can be pulled from any model in Geoanalytics or any layer in an OGC compliant web mapping service (WMS) stream.

The drawing and analysis tab contains controls for drawing polygons, points, paths, and icons which are shared with all users in real time. These can be annotated with information upon creation. Additionally, custom controls can be defined and associated with a user's role to provide in-depth analytical tools to different users.

The discussion tab contains a chat client for real-time chat with other users. In addition to this, the positions of participants are shown on the map, whether they are logged in by mobile client (tablet, phone) or to a desktop or laptop based client. This helps contextualize the interactions of participants and maintain situational awareness.

To facilitate the addition of formal data to the map, mobile web-forms can be developed to query participants using mobile devices for specific information about their location without requiring

them to be logged into the full client, which can be too heavyweight for users in hard-to-reach areas.

## Other applications

RENCI's Geoanalytics infrastructure has been / is being used in some capacity in the following projects:

- Gillings School of Public Health Farmers' Market Locator
- UNC School of Public Health global impact website
- NCB Prepared prototypes
- NOAA's collaborative agreement with RENCI on emergency management and situational awareness, dubbed the WxEM project.
- NARA's collaborative agreement with RENCI for scaling archiving cyberinfrastructure to billions of archival records, named CI-BER.

Geoanalytics is well-suited to projects in the environmental sciences, public health, and other projects requiring scalable cyberinfrastructure involving geography. It has been used to house and provide online access to such diverse data as:

- Pointwise data for public health applications.
- Very large and diverse archival metadata collections.
- NOAA weather forecast data.
- Census data.
- DOT road maps.
- Environmental modeling data, including SLOSH and ADCIRC.
- A complete set of 20m/pixel resolution LiDAR data for the state of NC.